UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A MULTIDIMENSIONAL APPROACH TO TEST ITEM SELECTION IN A PRACTICAL COLOR VISION TEST

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

NELDA JEAN MILBURN
Norman, Oklahoma
2002

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

A MULTIDIMENSIONAL APPROACH TO TEST ITEM SELECTION
IN A PRACTICAL COLOR VISION TEST

A Dissertation APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Robert Terry

Larry E. Toothaker

Shane Connelly

Jorge Mendoza

Michael R. Buckley

Acknowledgments

I would like to express my sincere gratitude to the members of my dissertation committee--to Drs. Terry, Mendoza, and Toothaker for instilling in me a desire to learn statistics, and to Drs. Connelly and Buckley for their willingness to serve on my committee even without knowing me as a student. I feel privileged to have been taught and mentored by the best. I would like to thank Dr. Robert Terry, my major professor, for believing in me. (To fully understand the significance of that statement, take a minute to mentally list the people who have honored you in a similar way, then take the time to find the words to thank them! My mother tops that list.) This research was possible because of the encouragement from Dr. David Schroeder and Dr. Henry Mertens, my supervisors and mentors at the Federal Aviation Administration Civil Aerospace Medical Institute who have supported my continued education in many ways.

I would be remiss if I failed to mention my comrade, Bill Farmer—we commiserated together and encouraged each other. I would like to recognize Dr. Dana Broach, Deborah Perry, Lena Dobbins, Kali Holcomb, and Carolyn Dollar for their contributions toward this research project. Dr. Uebersax was especially helpful and patient while I worked through the mechanics of LLCA. Finally, I would like my family to know that I appreciate all the times they made sacrifices so I could attend graduate school. Thank you, JaNena, JaNetta, JaNell, JaNiece, and Wesley for being perfect in every way!

Table of Contents

CHAPTER 4

CHAPTER 5

List of Tables

List of Photographs

List of Figures

Abstract

Previous research evaluated the color vision requirements and perceptual factors related to interpretation of color-coded weather radar information and resulted in an 80-item developmental test. The objective of this study was to use Item Response Theory (IRT) methods to select the most efficient set of those items without adversely affecting the reliability or item domain of the test. The resulting test was used to examine the effectiveness of using Located Latent Class Analysis (LLCA) to distinguish between individuals based on their capability of identifying colors on a CRT. The participants were 335 individuals with normal color vision and 200 with varying degrees of red-green types of color vision deficiency. BILOG-3 (with a 3-parameter IRT model specified) provided indices for discrimination, item difficulty, and guessing. The greater the discrimination value, the better the item discriminates high performers from low performers. Likewise, large difficulty values (in the positive direction) indicate more difficult items. Higher guessing values indicate a greater probability of guessing the correct choice. BILOG-3 also calculates an estimate of each individual's ability using the same scale as the item difficulty index. Using these indices, a composite scatterplot was constructed. With the item discrimination plotted on the Y-axis, and the item difficulty/ability estimate on the X-axis, boundary lines were drawn for the guessing parameter in addition to the minimum and maximum ability estimates for each anomaloscope diagnosis. Once the item's content was coded onto the scatterplot, items were evaluated on several dimensions to select the most efficient items. Maintaining the original passing criterion of no more than 1 error, high consistency was found between the original 80 items and the 50-item test ($K_{(172)}$=.95). Correlation between percent correct scores was also very high ($r_{(172)}$=.99). The original inter-item reliability (α=.96) and item content were unaltered by shortening the test. LLCA latent trait scores, CTT percent correct scores, and IRT ability estimates were highly correlated; however, agreement between LLCA latent trait class assignments and pass/fail on the CWRT was moderate, K (172) = .62 but not sufficiently high to be used as a methodology in safety-critical decisions. The unique, multidimensional approach to item selection using the IRT 3-parameter estimates while taking item content and ability range information into consideration yielded 50 highly reliable items.

CHAPTER 1

A Multidimensional Approach to Test Item Selection in a Practical Color Vision Test

Overview

The Federal Aviation Administration (FAA) requires all Air Traffic Control Specialists (ATCSs) applicants demonstrate their color discrimination ability because several ATCS tasks involve critical, non-redundant, color-coded information. Previous research (Mertens, 1990; Mertens & Milburn, 1998; Mertens, Milburn, & Collins, 1995, 2000) related to the development of 2 job-sample color vision tests demonstrated that some individuals who fail color vision screening tests may be able to perform en route and terminal option ATCS color tasks as well as people with normal color vision (NCV). The FAA has an immediate need for a concise, work-sample color vision screening test for Automated Flight Service Station (AFSS) ATCSs. Three experiments (Milburn & Mertens, 2002) were conducted examining numerous factors related to presenting such a color vision test on a CRT display. As a result of those experiments, using both participants with normal and deficient color vision, dichotomously scored performance data on 80 items exists. Errors were rare among participants with NCV and those with mild color abnormalities (Milburn & Mertens, 2002). Participants with severe color abnormalities made color confusions between virtually all colors. The poor performance of participants with red-green color vision deficiencies (CVD) supports the need for a color vision requirement for ATCSs at AFSS facilities that must use color-coded radar displays when briefing pilots on weather before and during flights. A detailed account of the development the Color Weather Radar Test (CWRT, Milburn & Mertens, 2002) also chronicled method refinements used to ensure reliable presentation of colors and to train test-takers.

Statement of the Problem

During the development of the CWRT, only classical test theory (CTT) methods were used to analyze the performance data, to select items, to compare performance between color palettes and between color vision groups. At the conclusion of each developmental phase, items were selected based on CTT methods for evaluation in subsequent experiments. For example, only items passed by *all* individuals with NCV were selected to serve as test trials in the next phase. However at each subsequent phase, a few people with NCV made an error or two on items previously passed by all participants in the NCV group. CTT did not allow the test developer to distinguish between random and systematic errors and also, within the structure of CTT, the test developer could not predict the performance of an individual participant. Furthermore, because CTT item parameters are sample dependent, performance on individual items could not be predicted across samples, hence performance on the items appeared inconsistent.

Objective

There are 3 major objectives addressed in this paper. The primary purpose was to create a more concise edition of the CWRT from the existing

80 dichotomously scored items by applying test development principles advocated by Item Response Theory (IRT) to overcome some of the difficulties resulting from having used CTT methods. Secondly, this study compared the efficiency of establishing pass/fail cut scores using ability estimates (derived from IRT) compared to total percent correct scores (obtained using CTT methods). Third, the data were examined using Located Latent Class Analysis (LLCA) to determine the extent to which latent classes can be modeled from perceptual color vision data acquired from individuals of varying types and degrees of CVD. The purpose of using LLCA was to evaluate the latent class assignments for potential pass/fail classifications as distinctions between those capable and incapable of the color identification tasks.

CHAPTER 2

Review of Test Theory

CTT (Gulliksen, 1950) has served as an esteemed guide to test development for several decades and continues to be heavily cited. Most readers and test developers are familiar with CTT but may not be familiar with modern test theory. For that reason, an overview of modern test theory method is presented. For a quick study, Embretson's (1996) article on the new rules of measurement helps bridge the gap between classical and modern test theories. For example, one old rule of CTT applies to test length—"longer tests are more reliable than shorter tests" (Embretson, 1996, p. 342). Embretson (1996) cites the Spearman-Brown prophesy formula (Brown, 1910; Spearman, 1910) and work by Guilford (1954) that demonstrate that lengthening a test allows true variance to increase more rapidly than error variance, resulting in a more reliable test. Lord and Novick (1968) give a specific example of the impact on reliability as a result of doubling the length of a test as an increase from an original reliability of .6 to an improved reliability of .75. Embretson (2000) demonstrates the same effect is present in the reverse—that is by shortening a test with a reliability of .86 by two-thirds of its original length results in an anticipated reliability of .80.

However, the new rule states that "shorter tests can be more reliable than longer tests" (Embretson, 1996, p. 342). By using adaptive testing to select items optimally appropriate to the person's ability level, the test can be shorter yet more precise, meaning it has less measurement error and is therefore more reliable. Embretson points out that the Spearman-Brown prophesy formula (Brown, 1910; Spearman, 1910) assumes that the test is lengthened with parallel parts and that is not the case using an adaptive test, because items are selected specifically based on the response pattern of the test-taker. Other ways that a shorter test can be more reliable than a longer test is to use modern test theory methods to select items that:

- Cover a range of difficulties that correspond to the range of abilities of the test-takers;
- Provide the most test information (item discriminability) across the range of abilities;
- Have a low probability of guessing the answer correctly;
- Are near the cut-point (if the cut-point is known); and
- Have low measurement error near the cut-point.

Embretson discusses 5 other old rules that have been replaced by new ones developed by modern test theory that are essentially captured in the list below.

Hambleton and Swaminathan (1985) discuss 5 shortcomings of CTT that they believe modern test theory overcomes.

1) Item difficulty and item discrimination depend on the particular examinee samples in which they were obtained.
2) Comparison of individuals on a particular ability (for example color discrimination ability) by a set of test items is possible only if the individuals are given the same or parallel test items.

3) Test reliability is directly related to variability and is usually defined in terms of parallel forms or re-presentation of the same test.
4) Within the CTT framework, a test administrator can not predict how an individual will perform on a test item, nor can a test designer adapt a test to match a particular examinee's ability level.
5) CTT presumes that measurement error variance is the same for all examinees.

Given the shortcomings of CTT, an overview of item response theory (IRT) is provided below with the rationale for its selection and use in the present experiment and a description of the fit of the model to color perception testing.

The beginnings of IRT can be traced back to the 1930's and 40's (Lawley, 1943, 1944; Richardson, 1936; and Tucker, 1946) and substantial theoretical contributions to the development of IRT have been made each decade since (Lazarsfeld, 1950; Lord, 1953a, 1953b; 1974a, 1974b, 1977, 1980, Birnbaum, 1957, 1958a, 1958b; Rasch, 1960, 1966a, 1966b; Mislevy & Bock, 1990). IRT has continued to gain acceptance and has been used in a variety of applications (Hambleton, 1983).

IRT is considered appropriate for selection of color weather radar test items from a large pool of items for three reasons:

1. Item characteristics are not group-dependent (Hambleton, Swaminathan, & Rogers, 1991).

2. The performance of a participant on each item can be predicted based on that individual's estimated ability.
3. The relationship between the participant's performance on a particular item and the individual's estimated ability can be described by a monotonically increasing function and graphed as an item characteristic curve (ICC). That function stipulates that as ability increases, the probability of a correct response to an item increases.

Currently, there are several commercially available computer programs to perform the item response analyses. BILOG-3 (Scientific Software Incorporated, Chicago, IL) is one such program that provides item analysis and test scoring for binary logistic models. BILOG-3 was used because it met the following model restrictions. First, each item was scored dichotomously as correct or incorrect. Second, BILOG-3 provides ability estimates *(θ, theta*) for each individual and provides difficulty (threshold) parameters for each item, while taking into account the probability of a correct response to a multiple-choice item as a result of guessing. Third, a general assumption of IRT is that *if* the examinee knows the correct answer to the item, the examinee will answer correctly. This is *not* the same as if the examinee gets the item correct, then he must have known the correct answer. In a multiple-choice test, there is always the likelihood of guessing the correct answer. (Guessing will be discussed in greater detail later.) Fourth, BILOG-3's underlying theoretical assumption of a monotonically increasing relationship between color vision ability and the probability of a correct response seems to fit the model. For example, as an individual's color vision ability improves (approaches what is considered "normal" color vision), the likelihood of a making a correct response on color identification tasks also increases based on previous research that examined accuracy as a function of diagnosis of degree of color vision ability.

A unique strategy using IRT item parameters (fully described in the Method section) was used to examine the 80 items of the CWRT with the purpose of reducing the overall length of the test while providing the most test information. Pass/fail performance on the resulting parsimonious test was compared to original pass/fail performance. The efficiency and consistency of the parsimonious test was evaluated. Additionally, IRT methods were used to evaluate individual performance and to establish a pass/fail criterion.

The next analysis involved evaluating the efficiency of Located Latent Class Analysis (LLCA) for determining class structure as a function of color vision ability (the latent class of interest). LLCA is a statistical method for identifying groups of related cases (latent classes) from multivariate categorical or dichotomous data. For example, it can be used to identify distinct diagnostic classifications given the absence or presence of certain symptoms. Most frequently it is used on attitudinal survey data to find types of structures in the responses. Likewise, it can be used to locate consumer sub-groups from demographic or preference surveys. The goal of LLCA is to classify cases of individuals into their most likely latent class. It is called latent class because the individual's class membership is not/or can not be directly observed by the information gathered from the manifest variables. However, the most basic latent class analysis methods are "limited by the restrictive assumption that all variables are [conditionally] independent within each latent class" (Uebersax, 1999, p. 283). The most basic assumption of conditional independence is met for the data in that one's response to a particular item does not alter the probability of other items in the test.

LLCA was used to identify an individual test-taker's latent class (color vision diagnostic category) using performance data obtained from the CWRT (a multi-item inventory designed to measure color vision ability). Individuals were classified as having normal or deficient color vision as determined by the Nagel Type I anomaloscope, which is considered the criterion test for making such determinations. Uebersax (1999) refers to these categories as disease-positive and disease-negative cases.

In addition to categorizing individuals as "normal or deficient," the anomaloscope further classifies individuals within the deficient category by type of deficiency (protan or deutan). Within each type category, there is a continuum ranging from individuals possessing superior color discrimination ability (e.g. those having a lower probability of misidentifying colored targets) to individuals with severe CVD (e.g. those having a higher probability of misidentifying the same colored target). Although there was considerable overlap of *symptoms* or *color identification behaviors* indicative of the two types of deficients, there were also type-specific behaviors. That had the potential to violate the conditional independence assumption of latent class analysis (LCA) because it implies that, "within each latent class, presence of one symptom is associated with a higher probability of presence of other symptoms. Thus, one would not necessarily expect a standard two-class LCA model to fit such data" (Uebersax, 1999, p. 286). Uebersax further explains that "the only way that LCA can accommodate conditional dependencies is to add spurious latent classes that are not truly present at the taxonic level. Not only does this inflate the number of supposed latent classes, but also the entire latent class solution becomes susceptible to distortion. Hence, the interpretation of any possibly non-spurious latent classes produced becomes suspect. If the goal is to identify groups that correspond in meaning and number to reality, then a method is required to account for association of items within latent classes." (Uebersax, 1999, p. 287). However, that was an issue for the analysis itself and interpretation of the results to determine whether color vision perception could be modeled using LCA. To do so, LCA must be robust to minor violations of the independence assumption and still lead to a solution that could model performance. The point was to determine whether it provided a more useful method for measuring performance than CTT or IRT.

What this may imply related to using LCA to classify individuals based on their performance on the CWRT is as follows:

(1) Because Protans and Deutans make both similar and different types of errors (response patterns), it is possible that unique class membership may be difficult to establish; and a third class may be

formed from those individuals whose behavior is similar (the overlap area).

(2) Alternatively, because multiple items with the same target color are presented, the probability of a color deficient failing all of those items is very high--probably only lessened by the likelihood of guessing an item correctly.

(3) Furthermore, because of underlying color-mixture principles (the way colors are made on the CRT by mixing red, green, and blue lights to create all colors) any color that is the result of a color mixture would also create related items. However, conditional independence is maintained for points 2 and 3 because a response to one item does not determine the response to another item.

Background

The previous discussion on test theory should provide the reader with an adequate review to follow the comparison of CTT, IRT, and LCA presented in the results. Although knowledge of the origin of the data may not be absolutely necessary to understand the relevance or significance of the findings, that information may explain the practical importance of the results and the cause of certain response patterns. Therefore, this review provides some background information on several diverse, yet relevant topics pertaining to development of the work-sample, color vision test for AFSS ATCSs. Technological changes within the AFSS substantially changed the controller's tasks and subsequently the color vision standards for ATCSs.

Access to radar imagery began in the mid-1960's and used a monochromatic (one color) coding scheme to display shades of gray representing 3 levels of weather intensity. By the early 1980's, color radar imagery was accessible from National Weather Service radars. Six intensity levels displayed in color replaced the gray coding scheme. The new digitized weather product had at least 3 design aspects that added a level of redundancy to interpreting the color-coding. Features such as zoom, level select, and blinking enhanced color perception by increasing size; hence protected the performance of controllers whether their color vision was normal or deficient.

Basically, 16 colors were possible on the CGA (color graphics adapter) monitors with the red, green, and blue color guns. Additional technological changes allowed computers to produce 256 colors and eventually 16.7 million colors. Simultaneously, the resolution on the displays improved from 400 x 200 to 1280 x 1024 pixels. The improved resolution allowed weather displays to be presented in greater detail that more closely represented weather patterns. By the mid-1990s, color weather radar product suppliers expanded from 7- to 16-levels to represent a finer gradation of weather intensities. These technological and product changes setup a moving target scenario in the development of a work-sample test. Not only were the job-tasks changing because of the product changes and improvements, but also presumably the level of color vision ability required in air traffic control personnel. Because the older system had redundant cues available along with color that the new systems did not, the new systems required the user to rely more heavily on the color-coded information for interpretation. For those reasons, confirmation of one's ability to interpret the color-coded information became necessary, thus leading to a color vision standard and subsequently the development of a color weather radar work sample test.

Color Vision Standards for All Controllers

Applicants for ATCS jobs were required to demonstrate their color vision discrimination ability as early as 1978 when Medical Guideline Letter Number B-5A-0002 was written. It required NCV for all ATCSs because a number of ATCS tasks involved critical, non-redundant, color-coded information (Adams & Tague, 1985; Lahey, Veres, Kuyk, Clark, & Smith, 1984, Lahey, Veres, Kuyk, & Clark 1984; Mertens, 1990; Mertens, Milburn, & Collins, 1996, 2000; Pickrel & Convey, 1983). Given the number of tasks that require the ATCSs to decipher the meaning of the color-coded material, it is not surprising that "normal color vision" was a pre-employment job requirement. Furthermore, the Equal Employment Opportunity Commission's (EEOC) regulation (29 C.F.R. Section 1613.705a) states that an agency may not make use of any selection test that screens out qualified handicapped persons unless it is shown to be job-related.

The first step was substantiating the need for a color vision requirement. Justification for the screening test was based on its job-relevance, especially the non-redundant nature of color-coding in some ATCS tasks. However, making the connection between the color vision screening test results and the on-the-job tasks was harder to establish when the screening test was not composed of actual ATCS materials and/or tasks. Ultimately, the determination of NCV and one's capability of performing the color-related tasks was based on passing a Pseudoisochromatic Plate (PIP)Test (fully described later).

The PIP tests were medical tests designed to measure genetically determined variations of color vision among individuals; however, they do not predict the potential satisfactory performance of air traffic control duties related to color vision. Therefore, the requirement in the standard that the applicant must have *normal* color vision should be interpreted to mean that the individual must be able to function normally in recognizing colors in the work environment. Consequently, ATCS applicants who did not pass the PIP tests should be given the opportunity to demonstrate their ability to recognize colors in the air traffic work environment.

Based on results from 2 studies (Mertens, 1990; and Mertens & Milburn, 1992) a new directive from the Federal Air Surgeon (Jordan, 1992 in MGLRM, 1995) was written that instructed regional flight surgeons to administer the Dvorine PIP to all individuals seeking initial employment with the FAA as ATCSs. In the event an applicant failed the Dvorine PIP then work sample tests were to be administered. At the time the MGLRM was written, only the practical tests for the en route (center) and terminal options existed. This report involves the development of an alternative color perception test for the flight service station option.

Normal Color Vision (NCV) Defined

At this point it is appropriate to provide a very brief and easy-to-understand discussion of CVDs and to define what constitutes NCV. In this context, NCV refers to the way approximately 95% of the population sees colors. Estimates vary, but about 8 to 10% of all males and about one-half of 1% of females see colors differently than the majority of people. Because a very high percentage of the population *match* colors in a similar way, they are said to have NCV. In contrast, some individuals are commonly called color blind; however, the more appropriate term is color deficient because the occurrence of total color blindness is very rare. The incidence of CVD occurs across ethnic groups with the prevalence varying but with Caucasian males showing the largest ratio of red-green color vision deficients. The deficiency can be acquired as a result of injury, drugs, or disease; but it is most commonly inherited as a result of a recessive trait carried on the X chromosome—which explains its more likely occurrence in males than females. Furthermore, there are types of CVD and the prevalence of each varies. Most people with CVD are said to have a red-green deficiency because they confuse colors that fall along a line between the colors red and green within the 2-dimensional chromaticity diagram. Which brings us to how colors are described. "Until 1931, the concept of color had no precise scientific basis and colors could be specified only by appeal to physical samples [such as a color wheel containing color swatches]. In that year, the International Commission on Illumination (CIE) adopted a system of color specification, which has lasted to the present time" (Kaiser & Boynton, 1996, p. 25). Colors were described in terms of x and y coordinates that could be plotted on a chart known as a chromaticity diagram. There are 3 points that should be made for a basic understanding of the data presented in this paper. Specifically: how are colors made with lights; what effect CVD has on the use of a CRT; and how is color vision measured?

Creating Colors with Pigments and Lights

First, understanding the fundamental differences between combining primary color pigments and combining colored lights to make secondary colors is essential. (The reader is directed to Mueller and Rudolph's (1966) primer on "Light and Vision" which contains an easy-to-read description of the distinction between processes.) For example, if red and green colored *pigments* such as paint are combined—the resulting color is almost black. In contrast, if red and green colored *lights* are combined in approximately equal proportions—the resulting color is yellow. Furthermore, if full intensity red, green, and blue lights are added together—the resulting color is white light. This can be demonstrated by using *Microsoft Paint* or a similar software product on a computer that allows adjustment of the amount of red, green, and blue to define custom colors. This is an important difference to understand about how colors are made with light (such as on a CRT) and pigments (such as a printed color vision test).

Measuring Color Vision

The next building block of information is an understanding of the human color receptors in the eye and why color deficient people confuse colors. In this study, only the red-green types of CVD are discussed, because people with blue-yellow deficiencies (called tritans) are rare and none participated in any of these experiments. Deficients have a reduced sensitivity to certain colors. Because the mixture of red and green creates a CRT yellow, people with CVD would confuse yellow with green if they have a reduced number of receptors for red (protan-type), or confuse yellow with red if they do not have sufficient receptors for green (deutan-type).

Given that discussion, how is color vision measured? The Nagel Type 1 anomaloscope was the primary instrument used to diagnose type and degree of red-green color deficiencies in this study. It was recommended by the National Research Council-National Academy of Sciences, (NRC-NAS) Committee on Vision (1981) for color vision testing. The anomaloscope capitalizes on the process of combining red and green lights to create yellow as discussed above. First, the anomaloscope uses a prism to split a beam of light into the colors of the spectrum (ROY G. BIV is the acronym that we learned in school) including red, orange, yellow, green, blue, indigo, and violet. The anomaloscope uses a mixture of spectral red and spectral green to match to a spectral yellow light. When the participant looks through the monocular instrument, an illuminated bipartite circle can be seen. The top half is a variable red-green mixture, and the bottom is spectral yellow. By varying the luminance of the spectral yellow in the bottom half of the circle and the proportion of red and green mixed in the top half of the circle, the participant matches the two halves of the circle. The participant is told to match both color and brightness. The instrument provides a measurement of the luminance and the amount of red and green used to produce a match between the two halves of the circle. People with NCV combine almost equal proportions of red and green to match to yellow, whereas proportions selected by people with CVD greatly vary. Several matches are made and the amount they vary in the proportions is their *anomaloscope matching range*. The scale was 0 (spectral green) to 73 (spectral red). People with NCV have a very small and consistent matching range; and, people with CVD, depending upon the severity of their deficiency, have much larger matching ranges. In fact, a person with a severe CVD may match pure spectral green (or red or both) to spectral yellow.

Although the anomaloscope targets were much larger than the smallest targets used, it follows that a relationship should exist between the anomaloscope diagnosis and performance on the CWRT because of the similarities inherent to the color mixtures of the two apparatus.

The reader is referred to the report (Milburn & Mertens, 2002) that documents the development of the work-sample test because it includes a detailed description of the apparatus, stimuli, materials, screening tests, practical tests, and performance measures relevant to the 3 experiments' data analyzed in this report. However, the following summary will probably suffice. Experiment 1 evaluated several factors pertinent to creating test stimuli such as optimal target size, selection of color palettes, and the presence or absence of a color legend. Classical test theory methods were used to select potential test trials from the large pool of items presented in Experiment 1. Experiment 2 involved refinement of the participant's task, incorporated findings from Experiment 1, evaluated test-retest reliability, and recommended a smaller set of items for use in Experiment 3. Experiment 3 evaluated 2 methods of responding to test trials and established a cut-score for the test.

Re-statement of the Experimental Objective

After such a lengthy review of the various facets of information relevant to this study, a re-statement of the purpose of this experiment seems pertinent. It was to create a more concise edition of the CWRT by selecting the best items from the existing 80 dichotomously scored items by applying test development principles advocated by IRT. Furthermore, this study compared the efficiency of establishing pass/fail cut scores using IRT ability estimates compared to total percent correct scores obtained using CTT methods. The final objective was to compare LLCA class membership to anomaloscope diagnosis, to IRT cut scores, and to CTT percent correct scores to evaluate the efficiency of the LLCA procedure for potential use in determining pass/fail status.

CHAPTER 3

Method

Participants

Prior approval for all procedures and use of human participants was obtained from the Institutional Review Board of the Civil Aerospace Medical Institute (CAMI). Volunteers were recruited and paid by an independent contractor of the Human Resources Research Division of CAMI. The informed consent of every participant was obtained prior to participation, and each participant was free to withdraw from the experiment without prejudice at any time.

All volunteers had at least 20/30 corrected visual acuity in both near and distant vision as determined with the Bausch and Lomb Orthorater.

This study analyzed existing data obtained from the participants in 3 experiments described on Table 1. Participants were 342 people with NCV (170 females and 172 males) with a mean age of 31.4 years (sd. = 10.2) and 204 people with varying types and degrees of CVD (24 females and 180 males) with a mean age of 37.4 years (sd. = 11.2). All participants were between 18 and 57 years of age. A complete breakdown of the number of participants in each color vision category in each experiment can be found in Table 1. The process for classifying the participant's color vision is fully described below.

Diagnostic Classification of Participants

Normal Trichromats (n = 342) comprise the majority of all individuals and have a high level of color discrimination ability. In summary, all normal trichromats made anomaloscope matches that fell between 33 and 48 on the anomaloscope scale, midpoint within plus or minus 2 standard deviations of the mean midpoint (M = 40.5), and whose matching range was less than 16 units. This classification also contains the normal trichromats that Pokomy, Smith, Verriest, and Pinckers called *deviant normal trichromats* and *weak normal trichromats* (1979). The anomaloscope color-matching behavior of normal trichromats varies by a relatively small amount, but it does vary. These subgroups of normal trichromats may be thought of as representing the *tails* of either the distribution for the matching range size (the *weak* normal trichromats) or the distribution for the matching range midpoint (the *deviant* normal trichromats). *Deviant or weak normal trichromats* may show a very slight reduction in color discrimination ability.

Simple anomalous trichromats (n = 92) are the mildest of inherited red-green CVD. This category includes (1) individuals whose mid-matching point falls more than 3 SD above or below the mean for normal trichromats, and whose matching range does not overlap the range of mean matches of normals, and (2) all individuals with a matching range greater than 15, but less than 26 scale units, even if their range of matches overlaps the means for normals. Those simple anomalous trichromats having a mean of matches above the mean for normal trichromats were classified as simple protanomalous (n = 28), and those with a mean of matches falling below the normal mean were classified as simple deuteranomalous trichromats (n = 64). Simple protanomalous or deuteranomalous trichromats may have mild to moderate impairment of color discrimination ability.

Extreme anomalous trichromats (n = 61) have severe impairment and were separated into extreme protanomalous (n = 25) and extreme deuteranomalous (n=36). The extreme anomalous individuals accept a wide range of matches, overlapping the range of matches accepted by both the normal trichromats and the simple anomalous trichromat. The extreme anomalous trichromats had a matching range greater than 25 and less than 73 scale units that frequently included the mean of the normal group as well as part of the simple deuteranomalous or protanomalous matching ranges. Individuals in the extreme protanomalous group typically had a midpoint of matches *above* the NCV group mean and also have a reduced sensitivity to long wavelength (red) light. Extreme deutranomalous individuals typically had a matching midpoint *below* the mean of normals and no evidence of a sensitivity loss to long wavelengths of light.

Dichromat Anomalous Trichromats (n = 51) were similarly separated into protan and deutan groups called protanopes (n = 28) and deuteranopes (n = 23), respectively. All dichromats have severe color deficiencies and protanopes, like the extreme protanomalous, have reduced sensitivity to long wavelengths whereas deuteranopes do not. Both protanopes and deuteranopes have a range of 73 scale units (i.e., they accept the entire range of possible matches on the anomaloscope). See Table 2 for a brief summarization and description of each category.

Materials

Dvorine PIP

The Second Edition (1953) Dvorine Pseudo-Isochromatic [sic] Plates (PIP) was administered to all participants. The Dvorine PIP achieved a high validity when compared to the Nagel Type I anomaloscope diagnosis of "normal trichromat" (Mertens & Milburn, 1993). The NRC-NAS Committee on Vision (1981) cited several reports (Belcher, Greenshields, & Wright, 1958; Frey, 1962; Sloan & Habel, 1956) that noted the high color vision screening validity of the Dvorine PIP.

For each plate of the Dvorine PIP, the task of the observer is to identify a multi-colored Arabic numeral(s) composed of three sizes of dots embedded in a multi-colored background of dots. The disguised number(s) differ only in color from the background dots; size and intensity remain constant; hence are "hidden" only from the color deficient observers. The Medical Guideline Letters Reference Manual (1995) for testing ATCS applicants with a failure criterion of more than 2 errors was used, which is consistent with diagnosis of NCV as defined by the Dvorine PIP.

Color Weather Radar Test (CWRT)

The CWRT is a work-sample color vision test for AFSS ATCSs that used archived weather maps with a legend composed of swatches of the colors used to non-redundantly color code 16 levels of weather. Each level is related to the increasing probability of turbulence, the intensity of precipitation, wind shear, lightning, and hail, which are factors related to the hazardous weather conditions of flight. Using the map legend composed of shades of blue, green, yellow, red, purple, and white, examinees must identify the color on the map designated by an arrow and then use the mouse to mark their responses. Photograph 1 is a sample test trial. The CWRT replicates 3 color palettes found in use at AFSSs.

Training was presented as self-paced, interactive PowerPoint slides with voice instructions that accompanied the automated demonstrations and was followed by 20 practice trials. The test does not assume any prior knowledge of color weather radar displays and requires only the most basic computer skills to test the participant's color vision ability. Items were coded as correct or incorrect (1, 0) based on a scoring method that required correct identification of the target by color category. The cut score for passing was set at no more than 1 error.

Procedure

Item Selection

The first step was to obtain item parameters for each of the 80 dichotomously coded items by using BILOG-3 with a 3-parameter model specified. BILOG-3's output included a discrimination index, item difficulties, and a guessing index, a-, b-, and c-parameters, respectively. The greater the a-value, the better the item discriminated high performers from low performers, the larger the b-value (in the positive direction), the more difficult the item, and the higher the c-value, the greater the probability the examinee had of guessing the correct choice.

BILOG-3 provided an ability estimate for all participants obtained separately for two response-modes administered to participants in Experiment 3. (Participants responded to each trial by using the mouse to mark their choice and in a separate administration, participants announced their choice using specified color names.) Using the ability estimates, descriptive statistics were calculated for each color vision type, degree of deficiency, and by anomaloscope diagnosis. See Table 10. Focusing on the computer-response mode data (because that response method was selected for use in the final version of the CWRT), item parameters were calculated and BILOG-3 item characteristic curves (ICCs) were examined individually. ICCs plot the a-value (item discrimination) on the Y-axis, and the b-value on the X-axis (item difficulty/theta, Θ). The height of the c-parameter on the Y-axis indicates the probability of a correct response to a multiple-choice item as a result of guessing. Additionally, BILOG-3 draws the item information curve on the same graph indicating the range of abilities for which that item provides information and the point at which the height of the curve is tallest (called the b-value), and it is the point at which the item provides the maximum information. The b-value position is also known as the inflection point on the ICC, that is the point at which the slope is highest. The item information curve is essentially the inverse of the measurement error, so a high, broad curve would indicate a wide range of abilities for which the measurement error is small. To further interpret the item information curve, the width on the X-axis describes the range of ability scores for which the item provides information and the height indicates the amount of information available as a function of the ability scale. For example, the first item plotted on Figure 1(a) is one that provides a limited amount of information (the dotted line) across a range of ability scores ranging from –3 to +1.5. In contrast, the item information curve (the dotted line) plotted on Figure 1(b) demonstrates an item that provides information *about* a smaller range of ability scores, -1.0 to +1.5, but provides a substantial *amount* of information near the zero point on the ability scale.

The ICC (the solid line) provides information about the probability of a correct response as a function of ability. For example, Figure 1(c) is the ICC for a very difficult item. Notice that the probability of a correct response is low and flat over a large range of ability scores and is equal to the c-parameter, or the probability of correctly guessing the correct answer. In contrast, Figure 1(d) is a plot of a very easy item (b= -3.53) indicating a .70 probability of a correct response given a low ability scale score of –3.0. Notice the high probabilities of a correct response across a range of low abilities (the solid line) and the very little test information available for the item shown in Figure 1(d). A composite scatterplot was constructed using the following steps.

1. With the a-values on the Y-axis, and b-values on the X-axis, a horizontal line was drawn to indicate the maximum c-value of all items. Descriptive statistics using IRT ability estimates were calculated for each diagnosis.

Boundaries, representing the ability range of each diagnostic color vision group, were drawn as vertical lines onto the composite scatterplot. See Table 4 for descriptive statistics and Figure 2 for a sample plot.

2. The item points were color-coded to represent each item's content (the target color). See Figure 3.
3. Items plotted to the left of the lowest ability score boundary line and items with a-parameters that fall below the c-value line were dropped. See Figure 4.
4. Items were then evaluated based on content, item discrimination, and item difficulty. For example, if several items with yellow targets appear in the same general location on the b-value scale (on the X-axis), they were considered items of similar difficulty and content. See Figure 5.
5. Since the goal of the CWRT is to select individuals who can perform *as well as people with NCV,* items were selected (to provide the greatest test information) from items plotted near that cut-point.

The scatterplot facilitated a multi-dimensional approach to selection of optimal items based on decisions related to a-, b-, and c- parameters, including ability cut-score range, and all as a function of item content.

CHAPTER 4

Results and Discussion

Using the described method for item selection, 50 items were chosen to meet the primary objective of the experiment—to create a more parsimonious test. Several analyses were conducted to answer the pertinent research questions concerning the efficiency and/or the appropriateness of the method used. The feasibility of using ability estimates rather than percent correct scores as a measure of performance and to establish cut-scores was examined. Generally speaking, this phase of the analyses involved a comparison of performance on the original 80 items with performance on the 50-item set and an evaluation of the consistency of 2 dependent variables (percent correct and ability estimates).

The first premise that was established was the effectiveness of the item selection process and its impact on the test's reliability. Because the Spearman-Brown prophesy formula predicted only a small possible improvement from α (172) = .964 to .971 by doubling the length of the original data set to 160 items, then the goal of shortening the test was to maintain the original test's high reliability. That goal was accomplished with the 50 selected items. The original inter-item reliability, α (172) = .964, was unaltered by shortening the test. Although no precedence was found in the literature for the multivariate approach to item selection using IRT's 3-parameter estimates coupled with item content and ability range information, the method yielded 50 highly reliable items. However, the 30 items that were not selected had a much lower Cronbach α (172) = .76. In addition, the correlation of percent correct was much lower for those 30 items than for the 50 items when compared to the total test score (r (172) = .77 vs. .99, p < .01, respectively). Taking into account the shorter length (30 items compared to 50 items) the Spearman-Brown prophesy formula predicted an improvement in reliability to .79 given an additional 30 parallel items.

The next step was to determine the effect shortening the test had on the distribution of scores.

Comparison of Errors on the 80- and 50-Item CWRT and the Dvorine PIP

To understand the distribution of scores and the extent of the overlap between the color vision groups, Table 5 presents a crosstabulation of errors on the Dvorine PIP compared with errors on the 80 item color weather radar test as a function of degree of deficiency. Scores for the NCV group fit tightly into a small range of errors both on the Dvorine and the CWRT, with no one making more than 2 errors on the Dvorine PIP or 3 errors the CWRT. In contrast, the scores for participants with mild to moderate deficiencies, (the simple category) are dispersed across a wide range of errors (from 0 to 12) on the Dvorine PIP coupled with a small number of errors, or in some cases, a perfect score on the CWRT. However, all of the dichromats (those with severe deficiencies) made between 10 and 13 errors on the Dvorine corresponding with 4 to 48 errors on the CWRT.

Table 6 is provided to show the relationship between the number of errors on the 50-item CWRT and errors on the Dvorine PIP. The most consequential effect of deleting 30 items was on those who made only a few errors on the 80-item test because of their proximity to the cut-score as will be discussed in the next section.

Establish Pass/Fail Criteria

It is generally held that establishing a pass criterion or cut-score for any test is a major concern of the test designer. Rarely does a dividing line naturally occur between novices and experts, between apprentice and journeymen, or between people with normal and defective color vision. As a result, an overlap exists between the two distributions of scores making it difficult to set a cut-score. Furthermore, there are no clear-cut guidelines presented in the literature to definitively establish a cut-score for each new test developed. However, several researchers familiar with selection tests (personal communications with D. Broach, June 2001; R.Terry, January 2000; & M. Pulat, October, 1995) have suggested establishing the criterion performance standard at the point that included the performance of 95% to 98% of the subject-matter-experts or journeyman level employees. That is, in this case, if 95% to 98% of all people with NCV perform at, or above, the criterion cut-score, it is probably accurate to say that the criterion is placed at an appropriate point to describe the way people with NCV perform. The cut-score should include all but the extreme tails of the distribution of people with NCV.

The goal was to establish a cut-score for passing the CWRT based on theory, empirical evidence, and safety considerations. Establishing the pass/fail criterion for a test is always important and is directly tied to the type and purpose of the test. For example, if a test were designed to select the people who were most knowledgeable of the topic of interest, the cut-point would be established based on pre-determined qualifications for the job at the entry level. Or, depending upon the job market, an organization could opt to select only the top 15 to 20% of all applicants. However, the purpose of *this* test is to evaluate individuals based on the color capabilities of people with NCV to determine whether an individual (applicant) possesses those same capabilities. Of all people with NCV 89.3% made zero errors, 5% made 1 error, 3.5% made 2 errors, and 1.7% made 3 errors on the CWRT in the computer response condition.

The cut-score was established in a previous study (Milburn & Mertens, 2002) at 2 standard deviations below the mean for people with NCV. The NCV group mean was 99.78% correct (sd = 0.716) that translated to only 1 error on the 80-item test (98.75% correct). A determination was made to err conservatively when setting the criterion because of the safety-critical aspect associated with the selection criterion. A contingency table of pass/fail performance of the 80-item CWRT with the Dvorine PIP, the initial screening test for AFSS ATCSs, is provided on Table 7. Previous studies (Mertens, 1990; Mertens & Milburn, 1992a, 1992b; Mertens, Milburn, & Collins, 1998) reported high predictive validity values for the Dvorine PIP test (all reporting Kappas near .90) for simulated ATCS color-identification tasks. As with the practical color vision tests for the en route and terminal options, a few people were able to perform the work-sample tasks but were unable to pass the Dvorine PIP screening test. Mertens (1990) reported false-positive rates ranging between 7.1% and 35.8% for the Dvorine PIP test for prediction of performance on the 7-level, color weather radar tests using small and large targets, respectively.

Concerning the 50-item CWRT, the decision was to maintain the cut score for passing at no more than 1 error. Approximately 95% of all people with NCV passed with that criterion.

Comparison Between the 80- and 50-item Tests

The objective was to ascertain the extent to which the item reduction altered the pass/fail performance of individuals while maintaining the same pass/fail criterion. Shortening the CWRT affected the pass/fail decision for a few participants. They were 2 NCV participants and 1 person with an extreme Deutan-type deficiency. The 3 people missed 2 items each on the 80-item test and as a result of dropping 30 items, they made only 1 error on the 50-item test. Each person had made an error on a different item that was dropped. See Table 8. Table 9 provides a crosstabulation of the pass/fail performance on the original 80-item test compared to the reduced item set. The analysis of the agreement between the 2 test scores indicated a Cohen's kappa value of K (172) = .946 and a correlation between the percent correct scores of r (172) = .993.

Comparison of Test Information Curves

Given the high correlation between percent correct scores and agreement between pass/fail performance, the next step was to produce test information curves for the original 80 items and then for the resulting 50-item test and visually compare the two curves. See Figure 6. The goal in test development is to produce a test item function that is high and flat over the range of thetas (Θ, ability estimates) where accurate measurement is desired. As a consequence of dropping items that provided test information for only very low-level ability people, the test information curve shows that the parsimonious test provides little information at the –3.0 ability level. Essentially, the curve is shifted to the right (in the positive direction) compared to the 80-item test. The most test information is available near the –1.0 level in the shorter test (Figure 6 Point A) compared to the highest point of the curve in the 80-item test (Figure 6, Point B) that was near the –2.5 point. An ability estimate near the –1.0 point is approaching the lower range of ability estimates for people with NCV as diagnosed by the anomaloscope. As discussed earlier, the area near the cut-point on the ability scale needs to be the strongest point of the test. To further explain, the majority of items that provide high test information need to be just below and just above the cut point on the item difficulty/ability scale to adequately sample that range of ability and do so with a high degree of accuracy. The strongest point of the parsimonious test (Point A) shifted nearer the cut-point (the minimum ability of the criterion group) compared to the position of the original test peak (Point

B). Both tests had very low measurement error near the cut point. The combination of these two factors provides evidence that the test provides accurate and consistent information near the cut point--some assurance for low false positive and false negative cut-point decisions.

Using IRT Ability Estimates as a Dependent Variable

To establish confidence in the IRT ability estimates for use as a dependent variable, its feasibility, reliability, and correlation to benchmark standards should be established. The ability estimate should demonstrate a high correlation with the anomaloscope-measured degree of deficiency and should be shown to be at least as consistent as the percent correct score in repeated measures. Furthermore, if the IRT invariance claim holds, the IRT ability estimates should be consistent across parallel forms. So, the first objective was to establish credibility of the measure by comparing the ability estimate to a criterion-referenced measure, the anomaloscope diagnosis.

Comparison of Ability Estimates to Degree of Deficiency

A strong relationship should exist between ability estimates (ranging from –3.998 to +0.514) and degree of deficiency (coded 0 – 4 representing NCV, simple, extreme, and dichromat degrees of deficiency) because both are measures of color discrimination ability. Figure 7 provides box plots to compare the ability estimates by anomaloscope diagnosis. Notice the relationship between the ability and the degree of deficiency and the small standard deviations for each group. A strong negative correlation between the ability estimate and the degree of severity of deficiency was found, r (172)

= - .81, meaning the more severe the CVD, the lower the color discrimination ability score. The correlation between percent correct on the CWRT and degree of deficiency was, as expected, also negative, r (172) = - .70, but smaller. See Figure 8.

Comparison of Percent Correct and Ability Estimates as Measures of Performance

In Experiment 2 of Milburn and Mertens' study (2002), a correlation between the percent correct for Time 1 and Time 2 was calculated to get an estimate of test-retest reliability, r (257) = .95. Therefore, a similar correlation was calculated to assess the relationship between the two test presentations using BILOG-3 individual ability estimates as the dependent variable. The objective was to determine how similar the two ability estimates were, and also to compare the correlations of the two dependent variables. The ability estimates calculated separately for Time 1 and Time 2 were very similar, r (257) = .987, and somewhat higher than found using percent correct scores.

The computer and verbal response conditions used in Experiment 3 provides another measure of reliability of the CWRT and the dependent variables. The close association between the 2 separate estimates of ability for the 2 response conditions graphed as a function of deficiency type, degree of deficiency, and by diagnosis on Figures 9, 10, and 11.

Reliability of Ability Estimates

The invariance property of IRT ability estimates was examined further by obtaining separate ability estimates derived from the dichotomously scored items as given in Experiments 1 and 2. Keep in mind that the selected items of the CWRT were part of a much larger test in those experiments. Furthermore, the number and distribution of participants within each color vision category were different for each experiment. However, unlike CTT statistics, IRT estimates are not sample specific. There are several comparisons possible for the ability estimates, (the person statistics), which can be made using both CTT and IRT methods. Adding or removing an item from the test can impact the observed percent correct scores using CTT, but IRT ability estimates are invariant if the same latent trait is measured.

Parallel Tests Reliability

Forty people participated in Experiment 1 and returned for Experiment 2 (6 people with NCV, and 34 people with CVD). Ability estimates were calculated independently for Experiment 1 and the repeated measures that are called Time 1 and Time 2 of Experiment 2. Pearson bi-variate correlations were calculated to compare the 3 ability estimates. The correlation between Experiment 1 and Time 1 of Experiment 2 was r (40) = .907, between Experiment 1 and Time 2 of Experiment 2 was r (40) = .914, and between Time 1 and Time 2 of Experiment 2 was r (40) = .988. These results are particularly interesting because (a) several task refinements were made between Experiments 1 and 2; (b) the composition of the item sets from which the ability estimates were calculated changed; and (c) Experiments 1 and 2 were about 1 year apart. The high correlations indicate a strong relationship exists between the two test scores. Furthermore, the correlation between Time 1 and Time 2 of Experiment 2, even for a small sample, indicates few differences between the two ability estimates.

Efficiency of Ability Estimates for Establishing Pass/Fail Criterion

Applying the same logic to set the cut-score for ability estimates as was used to establish the pass/fail criterion with percent correct scores described above, the cut-score was set at 2 standard deviations below the mean for people with NCV. The mean for the NCV group (Experiment 3) was a scale score of 0.415 with a standard deviation of 0.305. See Table 4. With the cut-score set to include 2 SD below the mean for the criterion group, the minimum passing scale score was –0.195.

A crosstabulation of pass/fail decisions (on the 80-item test) based on 2sd below the mean for the NCV group, k (172) = .811 determined from ability estimates and percent correct scores is shown on Table 10. The two methods agreed for 92.02% of the cases, but did not agree for 6 individuals with diagnosis as follows: 2 normals, 1 simple protan, 2 simple deutans, and 1 extreme deutan. Five of those individuals were failed using the ability estimate criterion but passed using the percent correct criterion. One simple deutan was passed based on the ability estimate and failed the percent correct criterion.

Another statistical method called Located Latent Class Analysis (LLCA) that used response patterns to determine item difficulty, assign a trait score, and to classify individuals was explored.

Located Latent Class Analysis (LLCA)

LLCA Computer Program

Based on personal communications with the LLCA programmer (J. Uebersax, March, 2002) analysis involving more than 20 to 25 items to estimate parameters is *unexplored*. Because the parsimonious CWRT contained 50 items, my goal was to analyze response patterns to all 50 items. Dr. Uebersax generously provided the source code and the program was re-dimensioned to allow for 50 items. Unfortunately, several problems were encountered while trying to estimate parameters with only 36 unique response patterns because of too few unique response patterns. Uebersax indicated that K+1 unique response patterns are needed to make estimates. Therefore, 20 of the most discriminating items were selected based on the same multidimensional graphing procedure previously described. The target colors of the 20 items were distributed as 5 each of yellow, green, and purple targets, 2 red, and 3 white targets. Table 11 provides the usual CTT information (such as mean percent correct, standard deviations, Pearson correlations with the total test score, and the resulting alpha if the item were deleted) for each of the selected items.

Using only participants from the computer response mode of Experiment 3 (see Table 1) 32 unique response patterns were discovered for 20 selected items of the CWRT because 137 individuals responded correctly to all 50 items. That was the goal of the test-developers—to create a test on which people with NCV consistently performed accurately. However, response patterns without variability do not provide any information about the relative difficulty of each item. Some explanation of the assumptions of IRT and LLCA must be addressed at this point. Both methods assume that the test measures a single latent ability or trait and that as an individual's ability increases, the likelihood of a correct response also increases. These assumptions are referred to as the unidimensional and monotonic assumptions. Traub (1983) pointed out that one must not conclude that a data set meets all of the assumptions of a model, but that a researcher should determine the adequacy of the fit of the data set to the model assumptions. When the fit is poor, the resulting calculations will be questionable. The purpose of interjecting this discussion at this point is to state that the model-data fit is unknown and to question whether the number of unique response patterns is adequate.

The LLCA results are organized as 4 main sections: analysis involving item parameters, measures of individual participant performance, classes or categories of participants followed by a comparison of CTT, IRT, and LLCA variables. Item parameters include item difficulties (from CTT) and threshold values (from BILOG and LLCA). Participant performance was measured by percent correct (an estimate of CTT's true score), latent trait ability estimates (theta values from BILOG), and latent trait scores (from LLCA). Three separate models were specified and the resulting LLCA class assignments were compared to anomaloscope diagnoses and CWRT pass/fail decisions.

Finally, comparative analyses to determine how well each of the methods classified individuals was conducted.

Item Analysis

Although selection of test trials were based on the IRT multi-dimensional approach previously described, it is possible to employ a selection strategy based on LLCA probabilities by selecting test items with the most polar probabilities. Table 12 lists the conditional rating probabilities for the simplest LLCA model with 2 latent classes and 2 manifest variables. A conditional probability is computed for each latent class given a correct response and given an incorrect response—resulting in 4 probabilities. The way this data were coded, the most discriminating items have conditional probabilities approaching 1 when the latent class number matches the manifest category number. For example, the latent class could be predicted from performance on a perfectly discriminating item 100% of the time because if an examinee responded incorrectly to the item (coded 1), it would mean that they belonged to latent class 1. Likewise, if they responded correctly (coded 2) to the item, then they belonged to latent class 2. Rarely in the real world do items such as that one exist; however, using LLCA conditional probabilities, the most discriminating items can be identified. For example, compare the conditional probabilities for item 1 to item 4 using Table 12. Both items report a high probability (.968 and .994) of membership in latent class 2 if the item is answered correctly (manifest category 2), but the likelihood of being in latent class 1 if the response is incorrect is much higher for item 1 than item 4 (.814 vs. .440). Creating a test by selecting items based on the described methodology should be addressed in a follow-on study to compare item selection strategies. However, the method that was used in this experiment was to *select* test trials based on IRT parameter estimates then use LLCA parameter estimates to evaluate the efficiency of the model for classifying participants based on performance on the selected items.

Reliability of Item Threshold Parameters

Recall that Experiment 3 (Milburn & Mertens, 2002) involved re-presentation of the same items in a unique random order using 2 separate response methods (computer mouse or orally, versions 1 and 2 respectively). Each version contained 32 unique response patterns, but they were not identical. Threshold parameters for the selected 20 items were calculated from the response patterns from each single presentation of the test and were compared to get a measure of reliability of the threshold parameters, r = .794. Table provides the slope, asymptote, and threshold parameters for each item for test version 1.

Using only the computer test version response patterns, the LLCA 20 item threshold values were compared to the CTT item difficulties and were very highly negatively correlated, r= -.967. (Recall that higher threshold values indicate more difficult items whereas the same CTT indicator is noted by lower values—hence an inverse relationship.)

The next analysis involved comparing the LLCA item threshold values with the IRT (BILOG-3) item thresholds and the resulting correlation was very high r=.955. The third comparison was between IRT item threshold values and the CTT item difficulties with a correlation of r= -.959.

For exploratory purposes only, the data sets for versions 1 and 2 were combined (58 unique response patterns) to determine whether increasing the number improved LLCA item threshold parameter estimates. When the LLCA item parameters calculated separately were compared to the combination, the correlations were very high for the 20 items, r=.931 and r=.950, indicating that for purposes of calculating item thresholds, few differences exist between the item parameter estimates calculated using 32 and 58 response patterns. This analysis does not attempt to pinpoint the number of responses needed to reliably obtain item thresholds, but was simply a gross check of the statistic. Presumably, some improvements to the item thresholds were possible with the added unique response patterns. That is a matter that could be examined using contrived data and multiple step-wise replications by adding additional unique response patterns until the optimum number is achieved.

LLCA Classification Efficiency

The LLCA program was executed specifying several different models using just the computer response version of the test. The theoretical rational for specifying 2, 3, and 4 latent classes relates to clinical methodology for color vision classification as 2 groups—normal/deficient, 3 color vision types—normals, protans, and deutans, and 4 degrees of color vision—

normal, mild deficients, moderate deficients, and severe deficients, respectively. The first model specified was the 2 latent class model (deficient/normal color vision—classes 1 and 2, respectively) with 2 manifest categories (incorrect/correct coded 1 and 2, respectively).

Two Latent Classes

The most general class assignment placed participants into two classes—deficient or normal color vision. Table 14 reports test sensitivity and specificity information. Sensitivity is the proportion of true positives that are correctly identified by the test and specificity is the proportion of true negatives that are correctly identified by the test. Based on this information from the sample studied, you could expect 31.67% of the participants with deficient color vision to indicate positive for the deficiency (class 1), while 100% of those with NCV would have normal test results (negative for the deficiency) or class assignments (class 2). Test sensitivity and specificity are known only if criterion test information is available (as was the case in our experiment) and is an approach used for quantifying the diagnostic ability of the test. Although, in practice and in most real world applications, such information is not available and we must rely on developmental or validation study results to determine how well the test predicts deficiency or color abnormality. Kappa was used to measure agreement between the LLCA 2-class assignments and the criterion test diagnosis of normal versus deficient color vision and agreement was poor (K (171) = .376. However, Kappa improved, K (172) =.722, when the simple-degree deficients were included with the normal group (because they tend to respond more like normals than deficients).

Three Latent Classes

Agreement between the LLCA 3-class assignments and the 3 anomaloscope type diagnoses was measured. Because Kappa is calculated by measuring agreement between like-coded classes and because LLCA class number assignments were unmapped to the anomaloscope class numbers, some deciphering was required. As with the 2-class analysis, group 2 was determined to be the normal group because individuals with all items correct were labeled group 2. Groups 1 and 3 were compared with both the deutan and protan groups. With protans coded as group 1 and deutans coded as group 3, agreement was poor, K (172)= .21, but improved somewhat when the coding was reversed, K (172)= .33. Table 15 shows the distribution between LLCA class assignments and the 3 criterion diagnostic type categories.

Four Latent Classes

Mapping between the 4 LLCA class assignments and the 4 degrees of deficiency was much easier to decipher because of the linear relationship between latent class location parameters (between -3 and +3) and the increasing degrees of deficiency. However, the normal group was still coded as group 2 with a latent class location parameter of +3 and the individual (rather than group of cases) with the most items coded as 1 (incorrect) had a latent class location parameter of –3 as in 3-class model analysis. Two additional groups were formed with intermediate location values of .3 and 1.207 with only 1 individual (a dichromat deutan) placed in the latter group. See Table 16. Naming the latent class groups produced by LLCA is similar to labeling the factors or constructs produced by factor analysis. At this point, little is known about the structure within the response patterns of this data set that would result in such a partitioning, but this exploratory work may serve as a reference in future studies. Agreement between the 4 class assignments and the diagnosis of degree of deficiency was weak, K (172)= .23. The SPSS coding for degree of deficiency was matched to the sequential latent class location parameters for purposes of the kappa analysis; however, not even mapping the normal group to class 2 was done with confidence because several from each diagnostic degree were also classified as group 2. Therefore, the meaning of membership in class 2 is unknown.

Comparison of LLCA Class Assignment to Pass/Fail on the CWRT

LLCA class assignments were compared to pass/fail decisions on the CWRT. First, the 2-class model results were tested and agreement was moderate, K (172) = .62 but not sufficiently high to be used as a methodology in safety-critical decisions. LLCA coded 18 individuals into class 2 who failed the CWRT. See Table 17.

Individual Participant Performance

In this section, 3 ways to compute a performance score for each participant were compared. CTT percent correct scores were previously compared to IRT ability estimates, but now both of those dependent variables will be compared to LLCA latent trait scores calculated from 3 specified models—2, 3, and 4-class models.

Comparison of IRT Ability Estimates to LLCA Latent Trait Scores

The IRT ability estimates were compared to the LLCA latent trait scores calculated under each model, r (172)=.847, .901, .939 and the correlation improved as the number of classes specified increased, perhaps the result of greater variability of latent trait scores assigned. Tables 18, 19, and 20 show the frequency and distribution of latent trait scores for each model tested. The latent trait scores ranged between –3 and +3 for all three models, but the *number* of unique scores assigned (8, 9, and 12) varied with the models (2-, 3-, and 4-classes, respectively). Tables 21, 22, and 23 present LLCA latent trait scores with the response patterns and frequencies, class assignments, and the probability of a class given a response as a result of fitting the 2-, 3-, and 4-class models.

Comparison of Ability Estimates with Percent Correct Scores

A previous analysis compared IRT ability estimates with percent correct scores to evaluate their efficiency as dependent variables but did not involve a direct comparison of the two scores to determine the extent of association. Using the original 80-item test, a correlation of r (172) = .903 was found between percent correct scores and the IRT ability estimate. The correlation improved considerably, (r (172) = .953, between the percent correct score and the ability estimate assessed from the 20 items, which is probably the result of eliminating random errors in one or both of the scores.

Comparison of LLCA Latent Trait Scores with Percent Correct

LLCA latent trait scores obtained under the 2-, 3-, and 4-class models were compared to the percent correct score on the original 80-item test. Correlations improved, r (172) = .826, .912, and .916, as the number of classes specified increased as was noted with the association between IRT ability estimates and LLCA latent trait scores previously discussed.

Classification Comparison

Finally, a comparative analysis evaluating the efficiency of each method's quantification of each individual's performance was conducted. Pass/fail cut scores were established at 2 standard deviations below the mean for the NCV group for each method's dependent variable. Kappas between CTT percent correct scores; LLCA latent trait scores; and IRT BILOG-3 ability estimates; and normal/deficient classifications made by the anomaloscope were K (172) = .528; .469; and .504, respectively.

Table 24 contains 4 contingency tables displaying the frequencies for the 3 measures reported above in addition to the LLCA class assignments using the 2-class model, K (172) = .376 previously discussed. Agreement with the anomaloscope diagnosis for each of the measures is low, yet very similar. However, if agreement were high between performance on the work-sample color vision test and a diagnostic test such as the anomaloscope, the work-sample test would not be necessary. But, the low agreement between actual performance and the criterion measure indicates that the two tests are not measuring the same level of performance required for the dichotomous classifications. Evidence of the overlap between the dichotomous classifications is apparent in the number of participants with mild to moderate CVD that capably perform the work-sample test.

CHAPTER 5

Conclusions

CTT item selection methods were used in the early stages of test development with the greatest drawback being CTT's inability to distinguish between random and systematic errors. To overcome that shortcoming, this study applied a unique strategy using IRT item parameters to examine the items of the CWRT with the purpose of shortening the test while providing the most test information. One advantage of IRT over CTT was that it placed participant performance and item performance on the same scale using one metric to describe the two most basic parts of testing and measurement. By linking the selection of items to the ability of a target population (in this case, people with NCV), it was possible to select individual items that predicted a high probability of a correct answer given NCV ability. Using CTT such an evaluation of individual items was not directly ascertainable.

The IRT selection method yielded 50 items and a pass/fail criterion was maintained at 98% correct that was established in previous research (Milburn & Mertens, 2002). The high Kappa, K (172) = .94, comparing the pass/fail performance on the original test (80 items) and the shortened version (50 items) indicated a high percentage of agreement between the two tests. Three people were reclassified from failing to passing as a result of dropping 30 items. They were 2 NCV participants and 1 person with a moderate Deutan-type deficiency. The 3 people missed 2 items each on the 80-item test and as a result of dropping 30 items, they made only 1 error on the 50-item test. Keep in mind that the process used to determine which items were dropped from the test was determined by the discriminability and difficulty of the items according a prescribed procedure and *did not* involve singling out items because people with NCV missed them. Therefore, the item reduction procedure should not adversely impact nor aid any particular color vision group more than any other.

The high correlations between the 80- and 50-item tests calculated separately for 2 dependent measures indicates few alterations in the total percent correct and ability estimates, but instead shows a shifting of the scale to fewer total items. Because Cronbach's index of internal consistency was so high, $\alpha(172) = .96$, a high correlation was expected between the parsimonious set of items and the original set.

The 50-item test retained the longer test's high reliability, α (172) = .96, and did so contrary to CTT canons that state that longer tests are more reliable than shorter tests. Furthermore, the retention of the high Cronbach alpha of the 50-item test was possible because the selection procedure eliminated less reliable items.

Limitations caused by too few unique response patterns in the CWRT data necessitated the selection of a smaller set of items in order to examine the feasibility of using LLCA to classify and grade participants. Using the same procedure as was used to select 50 items, 20 items were selected. LLCA latent trait scores and class assignments were compared to CTT percent correct scores, IRT ability estimates, anomaloscope diagnoses, and finally to pass/fail decisions on the CWRT. LLCA latent trait scores correlated very highly to CTT percent correct scores and IRT ability estimates and correlations improved with the number of classes in the model increased. Correlations ranged between .826 and .939. However, when the latent classes were dichotomized, the 2-class model agreement was moderate, K (172) = .62 when compared to pass/fail criteria decisions. Agreement of the latent class assignments not considered sufficiently high to be used as a methodology in safety-critical decisions. However, the scatterplot technique using the IRT item parameters and individual ability estimates to select test items by item content yielded 50 highly reliable items and should be considered over traditional CTT item selection methods.

References

Adams, A.J. & Tague, M.K. (1985). Performance of air traffic control tasks by protanopic color defectives. *American Journal of Optom. Physiology. Optics*. 62, 744-750.

Agoston, G.A. (1987). *Color theory and its application in art and design*. New York: Springer Verlag.

Belcher, S. J., Greenshields, K.W., and Wright, W.D. (1958). Colour vision survey using the Ishihara, Dvorine, Bostrom and Kugelberg, Bostrom, and American-Optical Hardy-Rand-Rittler tests. *British Journal of Ophthalmology* 42, 355-359.

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. Series Report No. 58-16. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Birnbaum, A. (1958a). On the estimation of mental ability. Series Report No.15. Project No. 7755-23, USAF School of Aviation Medicine,Randolph Air Force Base, Texas.

Birnbaum, A. (1958b). Further considerations of efficiency in tests of a mental ability. Series Report No.17. Project No. 7755-23, USAF School of Aviation Medicine,Randolph Air Force Base, Texas.

Brewer, C.A. (1994). Colour use guidelines for mapping and visualization. In MacEachren, A. and Taylor, D.R.F. (Eds.), *Visualization in modern cartography* (pp. 123-147). London, UK:Pergamon Press.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322.

Convey, J.J. (1985). Passing scores for the FAA ATCS color vision test: Report No. FAA-AM-85-7. Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment*. Vol. 8, 341-349.

Embretson, S.E. (2000). *Item Respnse Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Federal Aviation Administration *Guide for Aviation Medical Examiners* (1980, 1985, 1996, & 1999). Washington, DC: FAA Office of Aviation Medicine.

Frey, R.G. (1962). Die Trennscharfe einiger pseudo-isochromatischer Tafelproben. V Graefes *Arch Ophthalmol* 163:20-30 (as cited in NRC-NAS).

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

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

Hambleton, R. K. (Ed.) (1983). *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Swaminathan, H. (1985*). Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

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

Hunt, R.W.G. (1998). *Measuring Colour*, (3rd ed.). England: Fountain Press.

Kaiser, P.K. & Boynton, R.M. (1996). Human Color Vision (2nd ed.). Optical Society of America, Washington, DC.

Lahey, M.A., Veres III J.G., Kuyk, T.K., Clark, D.J., & Smith, E.N. (1984). Job analysis and determination of color vision requirements for air traffic control specialists. Report submitted under Contract OPM-50-83 with the U.S. Office of Personnel Management.

Lahey, M.A., Veres III J.G., Kuyk, T.K., & Clark, D.J. (1984). The impact of color weather radar equipment on the job of air traffic control specialists – flight service station option. Report submitted under Contract OPM-50-83 with the U.S. Office of Personnel Management.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburg*. 6, 273-287. (cited in Hambleton & Swaminathan, 1985).

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburg*, 62-A, 74-82. . (cited in Hambleton & Swaminathan, 1985).

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al., *Measurement and prediction*. Princeton: Princeton University Press. (cited in Hambleton & Swaminathan, 1985).

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika,* 18, 57-75.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement.* 13, 517-548.

Lord, F. M. (1974a). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39, 247-264.

Lord, F. M. (1974b). Quick estimates of the relative efficiency of two tests as a function of ability level. *Journal of Educational Measurement*, 11, 247-254.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Earlbaum.

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

Medical Guideline Letters Reference Manual, MGLRM (January, 1995). DOT/FAA, Washington, D.C: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W. (1990). Evaluation of functional color vision requirements and current color vision screening test for air traffic control specialists. Report DOT/FAA/AM-90/9. Washington, D.C: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W., & Milburn, N.J. (1992a). Performance of color-dependent tasks of air traffic control specialists as a function of type and degree of color vision deficiency. Report DOT/FAA/AM-92/28. Washington, D.C: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W., & Milburn, N.J. (1992b). Validity of clinical color vision tests for air traffic control specialists. Report DOT/FAA/AM-92/29. Washington, D.C.: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W., & Milburn, N.J. (1993). Validity of FAA-approved color vision tests for Class II and Class III aeromedical screening. Report DOT/FAA/AM-93/17. Washington, D.C: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W., Milburn, N.J., & Collins, W.E. (1995). Practical color vision tests for air traffic control applicants: En route center and terminal facilities. Report DOT/FAA/AM-95/13. Washington, D.C: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W., Milburn, N.J., & Collins, W. E. (1996). A further validation of the practical color vision test for en route air traffic control applicants. Report DOT/FAA/AM-96/22. Washington, D.C: Federal Aviation Administration, Office of Aviation Medicine.

Mertens, H.W., Milburn, N.J. (1998). Validity of clinical color vision tests for air traffic control. *Aviation, Space, and Environmental Medicine*, 69, 666-674.

Mertens, H.W., Milburn, N.J., & Collins, W. E. (2000). Practical color vision tests for air traffic control applicants: En route center and terminal facilities. *Aviation, Space, and Environmental Medicine,* 71, 1210-1217.

Milburn, N.J. & Mertens, H.W. (in press). Development of a work-sample color vision screening test for automated flight service station air traffic control specialist applicants. Washington, D.C: Federal Aviation Administration, Office of Aerospace Medicine.

Mislevy, R.J., & Bock, R. D. (1990). *BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items.* Chicago: International Educational Services.

Mueller, C.G., & Rudolph, M. (1966). *Light and vision.* Life Science Library, New York: Time, Inc.

Murch, G. (1984). The effective use of color: Perceptual principles. *Tekniques*, 8, 4-9.

National Research Council-National Academy of Sciences, Committee on Vision (NRC-NAS). (1981). *Procedures for testing color vision: Report of Working Group 41.* Washington, D.C: National Academy Press.

Pickrel, E. W., & Convey, J. J. (1983). Color perception and ATC job performance: Report FAA-AM-83-11. Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Pokorny, J., Smith, V. C., Verriest, G., & Pinckers, AJLG. (1979). *Congenital and acquired color vision defects*. New York: Grune & Stratton.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (cited in Hambleton & Swaminathan, 1985).

Rasch, G. (1966a). An item analysis which takes individual differences into account. British *Journal of Mathematical and Statistical Psychology*, 19, 49-57. (cited in Hambleton & Swaminathan, 1985).

Rasch, G. (1966b). An individualistic approach to item analysis. In P. Lazarsfeld, & N.B. Henry (Eds.), *Readings in Mathematical social science*. Chicago: Science Research Association, 89-107. (cited in Hambleton & Swaminathan, 1985).

Richardson, M.W. (1936). Relation between the difficulty and the differential validity of a test. *Psychometrika*. 1, 33-49. (cited in Lord & Novick, 1968).

Sloan L.L., & Habel A. (1956). Tests for color deficiency based on the pseudoisochromatic principle. *Arch Ophthalmology* 55, 229-239.

Smith, V.C. & Pokorny, J. (1977). Large-field trichromacy in protanopes and deuteranopes. *Journal of the Optical Society of America. 67(2)*, 213-220.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory.*

(pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika.* 11, 1-14. (cited in Lord & Novick, 1968).

Uebersax, J.S. (1999). Latent Class Analysis Frequently Asked Questions (FAQ). http://ourworld.comuserve.com/homepages/jsuebersax/faq.htm. Retrieved October 21, 2001.

Wyszechi, G., & Stiles, W.S. (1982). *Color Science* (2nd ed.). New York: Wiley.

Appendix A

Tables

Table 1

Participants in each Experiment by Gender and Color Vision Group

	Diagnosis							
Experiment		Simple		Extreme		Dichromat		
	Normal	Protan	Deutan	Protan	Deutan	Protan	Deutan	Total
Male	27	7	14	8	14	8	8	86
Female	26		3		1		1	31
EXP 1	**53**	**7**	**17**	**8**	**15**	**8**	**9**	**117**
Male	94	13	19	11	11	10	8	166
Female	83	1	6	1				91
EXP 2	**177**	**14**	**25**	**12**	**11**	**10**	**8**	**257**
Male	51	6	13	5	10	9	6	100
Female	61	1	9			1		72
EXP 3	**112**	**7**	**22**	**5**	**10**	**10**	**6**	**172**
Total	**342**	**28**	**64**	**25**	**36**	**28**	**23**	**546**

Table 2

Anomaloscope Classification

Code	Diagnosis	Rationale/Method
N	Normal [a]	Midpoint of color matches between 33 and 48 with a range less than 16 units.
SP	Simple Protan	Midpoint of color matches greater than 40.5 with a range of 16 to 25 units or midpoint greater than 48 and range less than 16 units.
EP	Extreme Protan	Color matching range of 25 to 72 units with a systematic decrease in matching brightness as color approaches red.
DP	Dichromat Protan	Color matching range of 73 units with a systematic decrease in matching brightness as color approaches red.
SD	Simple Deutan	Midpoint of color matches less than 40.5 with a range of 15 to 25 units and little variation in matching brightness or midpoint less than 33 and range less than 16 units.
ED	Extreme Deutan	Color matching range of 25 to 72 units and little variation in matching brightness.
DD	Dichromat Deutan	Color matching range of 73 units with little variation in matching brightness.

[a] includes weak or deviant normal trichromat

Table 3

Descriptive Statistics for the Verbal and Computer Response Conditions as a Function of Anomaloscope Diagnosis

Diagnosis	Response Condition	N	Minimum	Maximum	Mean	SD
Normal	Verbal	108	93.75	100	99.42	1.189
	Computer	112	96.25	100	99.78	0.716
Simple Protan	Verbal	7	93.75	100	93.75	2.28
	Computer	7	97.50	100	97.50	1.25
Simple Deutan	Verbal	21	87.50	100	98.87	2.76
	Computer	22	98.75	100	99.89	0.37
Extreme Protan	Verbal	5	53.75	98.75	82.00	19.79
	Computer	5	67.50	97.50	84.50	15.17
Extreme Deutan	Verbal	10	52.50	100	91.13	14.76
	Computer	10	65.00	100	91.37	11.40
Dichromat Protan	Verbal	10	61.25	88.75	78.25	9.63
	Computer	10	63.75	92.50	79.75	10.39
Dichromat Deutan	Verbal	6	53.75	90.0	79.37	14.18
	Computer	6	40.00	95.00	74.79	20.11

Table 4

Descriptive Statistics for the BILOG-3 Ability Estimates for the Computer Response Mode

Diagnosis	N	Minimum	Maximum	Mean	SD
Normal	112	-0.879	0.514	0.415	0.305
Simple Protan	7	-0.907	0.514	-0.227	0.704
Simple Deutan	22	-0.461	0.514	0.442	0.239
Extreme Protan	5	-3.011	-0.907	-1.838	0.933
Extreme Deutan	10	-3.012	0.514	-0.887	1.234
Dichromat Protan	10	-2.985	-1.499	-2.235	0.523
Dichromat Deutan	6	-3.998	-1.186	-2.379	1.006

Table 5

Crosstabulation of Errors on the 80-Item CWRT by Number of Errors on the Dvorine PIP Reported by Degree of Deficiency

		Total Errors on the Dvorine PIP											
		0	1	2	3	6	8	9	10	11	12	13	Total
Degree	CWRT												
Normal	0	84	13	2									99
	1	6											6
	2	4											4
	3	1	1										2
Simple	0	11	4	2	2	3				1			23
	1							1	1		1		3
	2	1							2				3
Extreme	0		1				1				1		3
	1										1		1
	2									2	1		3
	4									1			1
	7								1				1
	8							1					1
	10										1		1
	16									1			1

	25								1				1
	26								1				1
	28								1				1
Dichromat	4											1	1
	6									1			1
	7								1	1			2
	9								1	1			2
	11								1				1
	13									1			1
	14										1		1
	16								1				1
	18										1		1
	19									1			1
	24									1			1
	27										1		1
	29								1	1			2
	48									1			1
TOTAL		1	19	4	2	3	1	2	12	13	8	1	172

Table 6

Crosstabulation of Errors on the 50-Item CWRT by Number of Errors on the Dvorine PIP Reported by Degree of Deficiency

		Total Errors on the Dvorine PIP											
		0	1	2	3	6	8	9	10	11	12	13	Total
Degree of Deficiency	CWRT Errors												
Normal	0	91	13	2									106
	1	1	1										2
	2	3											3
Simple	0	11	4	2	2	3		1		1			24
	1	1							2		1		4
	2								1				1
Extreme	0		1				1				2		4
	1										1		1
	2									2			2
	3							1					1
	4									1			1
	7								1				1
	8										1		1
	12									1			1

	18								1				1
	20								2				2
Dichromat	4											1	1
	5								1	1			2
	6									1			1
	7								1				1
	10								1				1
	11									1			1
	12								1				1
	13										2		2
	16									1			1
	18									1			1
	22										1		1
	24									1			1
	25								1				1
	35									1			1
TOTAL		10 7	19	4	2	3	1	2	12	12	8	1	171

Table 7

Crosstabulation of Pass/Fail Performance on the 80-item Test and the Dvorine PIP

Distribution	
Simple Protan	2
Extreme Protan	5
Extreme Deutan	6
Dichromat Protan	10
Dichromat Deutan	6

Distribution	
Normal	6
Simple Protan	1

		Dvorine	
		Fail	Pass
80-Items	Fail	29	7
	Pass	12	123

Distribution	
Simple Protan	1
Simple Deutan	8
Extreme Deutan	3

Distribution	
Normal	105
Simple Protan	3
Simple Deutan	14
Extreme Deutan	1

Table 8

Crosstabulation of Pass/Fail Performance on the 50- Item CWRT and the Dvorine PIP

		Dvorine	
		Fail	Pass
CWRT	Fail	27	3
	Pass	14	127

Distribution	
Simple Protan	1
Extreme Protan	5
Extreme Deutan	5
Dichromat Protan	10
Dichromat Deutan	6

Distribution	
Normal	3

Distribution	
Simple Protan	2
Simple Deutan	8
Extreme Deutan	4

Distribution	
Normal	108
Simple Protan	4
Simple Deutan	14
Extreme Deutan	1

Table 9

Crosstabulation of Pass/Fail Performance on the Original and Reduced Item Sets

		Original Item Set	
		Pass	Fail
Reduced Item Set	Pass	136	3
	Fail	0	33

Table 10

Crosstabulation of Pass/Fail Decisions Using 2 sd Below the Mean for Normals with Ability Estimates and Percent Correct Scores Calculated from the Computer Response Mode of Experiment 3 (80-item Test)

		Dependent Variable Ability Estimate		
Diagnosis	Dependent Variable Percent Correct	Pass	Fail	Total
Normal	Pass	98	2	100
	Fail	6	6	12
Simple Protan	Pass	3	1	4
	Fail		3	3
Simple Deutan	Pass	20	1	21
	Fail	1		1
Extreme Protan	Pass			
	Fail		5	5
Extreme Deutan	Pass	3	1	4
	Fail		6	6
Dichromat Protan	Pass			
	Fail		10	10
Dichromat Deutan	Pass			
	Fail		6	6
Total	Pass	124	5	129
	Fail	7	36	43

Table 11

Classical Test Theory Information[α]

	Item	Mean percent	Standard	Pearson	Alpha if
	Number	Correct	Deviation	Correlation	Item
					Deleted
1	4	.884	.3215	.652	.9104
2	9	.959	.1982	.468	.9141
3	17	.895	.3070	.791	.9051
4	19	.948	.2233	.610	.9103
5	15	.988	.1075	.489	.9139
6	23	.971	.1685	.569	.9116
7	26	.971	.1685	.583	.9113
8	37	.983	.1313	.367	.9152
9	40	.924	.2651	.713	.9075
10	42	.971	.1685	.612	.9108
11	47	.959	.1982	.622	.9102
12	43	.977	.1512	.522	.9127
13	45	.965	.1840	.694	.9089
14	46	.977	.1512	.616	.9111
15	54	.942	.2347	.604	.9105
16	62	.971	.1685	.455	.9137
17	70	.971	.1685	.498	.9129
18	71	.983	.1313	.617	.9116
19	80	.878	.3283	.657	.9104
20	84	.924	.2651	.512	.9134

[α] alpha = .9154, standardized item alpha = .9217

Table 12

Conditional Rating Probabilities

Item	Latent Class	Manifest Category	Conditional Probability
1	1	1	0.8144
	1	2	0.1856
	2	1	0.0319
	2	2	0.9681
2	1	1	0.3364
	1	2	0.6636
	2	1	0.0038
	2	2	0.9962
3	1	1	0.7361
	1	2	0.2639
	2	1	0.0205
	2	2	0.9795
4	1	1	0.4407
	1	2	0.5593
	2	1	0.0059
	2	2	0.9941
5	1	1	0.0577
	1	2	0.9423
	2	1	0.0005
	2	2	0.9995
6	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978
7	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978
8	1	1	0.0000
	1	2	1.0000
	2	1	0.0000
	2	2	1.0000

Table 12 (continued). Conditional Rating Probabilities

Item	Latent Class	Manifest Category	Conditional Probability
9	1	1	0.5838
	1	2	0.4162
	2	1	0.0104
	2	2	0.9896
10	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978
11	1	1	0.3364
	1	2	0.6636
	2	1	0.0038
	2	2	0.9962
12	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978
13	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978
14	1	1	0.1715
	1	2	0.8285
	2	1	0.0016
	2	2	0.9984
15	1	1	0.4407
	1	2	0.5593
	2	1	0.0059
	2	2	0.9941
16	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978

Table 12 (continued). Conditional Rating Probabilities

Item	Latent Class	Manifest Category	Conditional Probability
17	1	1	0.2273
	1	2	0.7727
	2	1	0.0022
	2	2	0.9978
18	1	1	0.1715
	1	2	0.8285
	2	1	0.0016
	2	2	0.9984
19	1	1	0.8340
	1	2	0.1660
	2	1	0.0364
	2	2	0.9636
20	1	1	0.6266
	1	2	0.3734
	2	1	0.0125
	2	2	0.9875

Table 13

Bilog-3 Rescaled Item Parameters for Computer Response Mode

Item	Slope	Threshold	Asymptote
1	1.550	-1.595	0.157
2	0.814	-3.197	0.164
3	1.741	-1.791	0.112
4	1.088	-2.672	0.155
5	1.136	-3.900	0.163
6	1.350	-3.036	0.165
7	1.151	-3.170	0.158
8	1.029	-3.593	0.178
9	1.547	-2.039	0.170
10	1.156	-3.170	0.158
11	1.346	-2.748	0.165
12	1.064	-3.405	0.164
13	1.552	-2.872	0.152
14	1.334	-3.258	0.157
15	1.018	-2.603	0.159
16	0.866	-3.445	0.164
17	0.848	-3.493	0.161
18	1.264	-3.510	0.159
19	1.835	-1.474	0.160
20	1.840	-1.768	0.249

Table 14

LLCA 2-Class Assignment as a Function of Normal/Deficient Anomaloscope Diagnosis

		Anomaloscope Diagnosis		
	Class	Normal	Deficient	Total
LLCA 2-Class Assignment	1		19	19
	2	112	41	153
	Total	112	60	172

Table 15

LLCA 3-Class Assignment as a Function of Anomaloscope Diagnosis of Type of Color Vision Deficiency

		Anomaloscope Diagnosis			
	Class	Protan	Normal	Deutan	Total
LLCA 3-Class Assignment	1			1	1
	2	10	112	31	153
	3	12		6	18
	Total	22	112	38	172

Table 16

LLCA 4-Class Assignment as a Function of Anomaloscope Diagnosis of Degree of Color Vision Deficiency

		Anomaloscope Degree Diagnosis				
	Class	Normal	Simple	Extreme	Dichromat	Total
LLCA 4-Class Assignment	1				1	1
	2	112	10	29	2	153
	3		5		12	17
	4				1	1
	Total	112	15	29	16	172

Table 17

LLCA 2-Class Assignment for CWRT Pass/Fail Decision

		CWRT Pass/Fail Decision		
	Class	Fail	Pass	Total
LLCA 2-Class Assignment	1	19		19
	2	18	135	153
	Total	37	135	172

Table 18

LLCA Latent Trait Scores [α] *Resulting From Fitting a Two-Class Model*

Latent Trait Score	Observed	Percent
-3.00000	9	5.2
-2.99990	3	1.7
-2.99160	3	1.7
-2.05540	1	.6
.00000	3	1.7
2.76800	7	4.1
2.99820	9	5.2
3.00000	137	79.7

N = 172 [α]mean = 2.39 SD = 1.76

Table 19

LLCA Latent Trait Scores $^{\alpha}$ Resulting From Fitting a Three-Class Model

Latent Trait Score	Observed	Percent
-3.00000	1	.6
.00000	3	1.7
.29920	1	.6
.30000	10	5.8
.30160	3	1.7
.45810	1	.6
2.64470	7	4.1
2.99620	9	5.2
3.00000	137	79.7

N = 172 $^{\alpha}$mean = 2.66 SD = .95

Table 20

LLCA Latent Trait Scores ^{a} Resulting From Fitting a Four-Class Model

Latent Trait Score	Observed	Percent
-3.00000	1	.6
.00000	3	1.7
.30000	3	1.7
.30020	2	1.2
.30100	1	.6
.30560	2	1.2
.33010	3	1.7
.44430	3	1.7
.80030	1	.6
1.96680	7	4.1
2.93860	9	5.2
2.99790	137	79.7

N = 172 ^{a}mean = 2.64 SD = .94

Table 21

CWRT Data and Results of Fitting a Two-Class Model

Response Pattern[a]	Observed	Expected	Class	*P* (class|response)	Latent Trait Score
22222222222222222222	137	133.70	2	1.0000	3.0000
12112112121212222222	1	0.00	1	1.0000	-3.0000
12222222222222122222	1	0.03	2	0.9613	2.7680
12112222122222222211	1	0.03	1	1.0000	-3.0000
12122222122222222212	1	0.03	1	0.9986	-2.9916
22222222222222222212	2	5.05	2	0.9997	2.9982
21222222222222222222	1	0.51	2	0.9997	2.9982
12112222122121222211	1	0.00	1	1.0000	-3.0000
11222222122222112222	1	0.00	1	1.0000	-2.9999
21111221212211222212	1	0.00	1	0.0000	0.0000
12222222222222222222	2	4.41	2	0.9997	2.9982
12222222222222222212	2	0.17	2	0.9613	2.7680
11111112111111111112	1	0.00	1	1.0000	-3.0000
12122221122222111211	1	0.00	1	0.0000	0.0000
22212222221222221211	1	0.00	1	1.0000	-2.9999

Table 21 (continued). CWRT Data and Results of Fitting a Two-Class Model

Response Pattern[a]	Observed	Expected	Class	*P* (class|response)	Latent Trait Score
12122222122122222211	1	0.01	1	1.0000	-3.0000
12122111121212122222	1	0.00	1	0.0000	0.0000
11122122221212212211	1	0.00	1	1.0000	-3.0000
12122222222222122211	1	0.02	1	1.0000	-2.9999
22222222222222222211	2	0.07	2	0.9613	2.7680
21112122221222122212	1	0.00	1	1.0000	-3.0000
22122222122222222211	1	0.01	1	0.9986	-2.9916
12122222111122122111	1	0.00	1	1.0000	-3.0000
12112212112211121111	1	0.00	1	1.0000	-3.0000
12122222222222222222	1	0.10	2	0.9613	2.7680
22122212212222222212	1	0.00	1	0.9986	-2.9916
21222222222222212222	1	0.00	2	0.9613	2.7680
22212222222222222222	1	0.79	2	0.9997	2.9982
22222222222222122222	1	0.79	2	0.9997	2.9982
22222222222222221222	1	0.30	2	0.9997	2.9982
12222222122222222221	1	0.00	1	0.8426	-2.0554
22122222222222222222	1	2.80	2	0.9997	2.9982

[a] 1 = fail, 2 = pass

Table 22

CWRT Data and Results of Fitting a Three-Class Model

Response Pattern[a]	Observed	Expected	Class	*P* (class|response)	Latent Trait Score
22222222222222222222	137	131.89	2	1.0000	3.0000
12112112121212222222	1	0.00	3	1.0000	0.3000
12222222222222122222	1	0.03	2	0.8684	2.6447
12112222122222222211	1	0.05	3	1.0000	0.3000
12122222122222222212	1	0.06	3	0.9994	0.3016
22222222222222222212	2	4.84	2	0.9986	2.9962
21222222222222222222	1	0.48	2	0.9986	2.9962
12112222122121222211	1	0.00	3	1.0000	0.3000
11222222122222112222	1	0.00	3	1.0000	0.3000
21111221212211222212	1	0.00	3	0.0000	0.0000
12222222222222222222	2	4.24	2	0.9986	2.9962
12222222222222222212	2	0.18	2	0.8684	2.6447
11111112111111111112	1	0.01	1	1.0000	-3.0000
12122221122222111211	1	0.00	3	0.0000	0.0000
22212222221222221211	1	0.00	3	1.0000	0.3000
12122222122122222211	1	0.02	3	1.0000	0.3000
12122111121212122222	1	0.00	3	0.0000	0.0000

Table 22 (continued). CWRT Data and Results of Fitting a Three-Class Model

Response Pattern[a]	Observed	Expected	Class	*P* (class|response)	Latent Trait Score
11122122221212212211	1	0.00	3	1.0000	0.3000
12122222222222122211	1	0.04	3	1.0000	0.3000
22222222222222222211	2	0.07	2	0.8684	2.6447
21112122221222122212	1	0.00	3	1.0000	0.3000
22122222122222222211	1	0.02	3	0.9994	0.3016
12122222111122122111	1	0.00	3	1.0000	0.3000
12112212112211121111	1	0.00	3	0.9998	0.2992
12122222222222222222	1	0.10	2	0.8684	2.6447
22122212212222222212	1	0.00	3	0.9994	0.3016
21222222222222212222	1	0.00	2	0.8684	2.6447
22212222222222222222	1	0.78	2	0.9986	2.9962
22222222222222122222	1	0.78	2	0.9986	2.9962
22222222222222221222	1	0.26	2	0.9986	2.9962
12222222122222222221	1	0.01	3	0.9415	0.4581
22122222222222222222	1	2.74	2	0.9986	2.9962

[a] 1 = fail, 2 = pass

Table 23

CWRT Data and Results of Fitting a Four-Class Model

Response Pattern[a]	Observed	Expected	Class	*P* (class|response)	Latent Trait Score
22222222222222222222	137	135.40	2	0.9988	2.9979
12112112121212222222	1	0.00	3	0.9988	0.3002
12222222222222122222	1	0.03	2	0.4644	1.9668
12112222122222222211	1	0.05	3	0.9938	0.3056
12122222122222222212	1	0.06	3	0.8412	0.4443
22222222222222222212	2	3.93	2	0.9663	2.9386
21222222222222222222	1	0.37	2	0.9663	2.9386
12112222122121222211	1	0.00	3	0.9988	0.3002
11222222122222112222	1	0.00	3	0.9668	0.3301
21111221212211222212	1	0.00	3	0.0000	0.0000
12222222222222222222	2	3.41	2	0.9663	2.9386
12222222222222222212	2	0.20	2	0.4644	1.9668
11111112111111111112	1	0.00	1	1.0000	-3.0000
12122221122222111211	1	0.00	3	0.0000	0.0000
22212222221222221211	1	0.00	3	0.9668	0.3301
12122222122122222211	1	0.02	3	0.9938	0.3056
12122111121212122222	1	0.00	3	0.0000	0.0000

Table 23 (continued). CWRT Data and Results of Fitting a Four-Class Model

Response Pattern[a]	Observed	Expected	Class	*P* (class|response)	Latent Trait Score
11122122221212212211	1	0.00	3	1.0000	0.3000
12122222222222122211	1	0.04	3	0.9668	0.3301
22222222222222222211	2	0.08	2	0.4644	1.9668
21112122221222122212	1	0.00	3	0.9989	0.3010
22122222122222222211	1	0.02	3	0.8412	0.4443
12122222111122122111	1	0.00	3	1.0000	0.3000
12112212112211121111	1	0.00	3	1.0000	0.3000
12122222222222222222	1	0.11	2	0.4644	1.9668
22122212212222222212	1	0.00	3	0.8412	0.4443
21222222222222212222	1	0.00	2	0.4644	1.9668
22212222222222222222	1	0.60	2	0.9663	2.9386
22222222222222122222	1	0.60	2	0.9663	2.9386
22222222222222221222	1	0.19	2	0.9663	2.9386
12222222122222222221	1	0.01	4	0.4996	0.8003
22122222222222222222	1	2.16	2	0.9663	2.9386

[a] 1 = fail, 2 = pass

Table 24

Comparison of CTT, IRT, and LLCA Pass/Fail Decisions as a Function of Normal or Deficient Anomaloscope Diagnoses[a]

		Anomaloscope Diagnosis	
		Normal	Deficient
CTT Percent Correct 2 SD below Mean of Normal Group	Pass	105	27
	Fail	7	33

K(172) = .528

		Anomaloscope Diagnosis	
		Normal	Deficient
IRT Ability Estimate 2 SD Below Mean of Normal Group	Pass	107	30
	Fail	5	30

K(172) = .504

		Anomaloscope Diagnosis	
		Normal	Deficient
LLCA Latent Trait Score[b] 2 SD below Mean of Normal Group	Pass	111	35
	Fail	1	25

K(172) = .469

		Anomaloscope Diagnosis	
		Normal	Deficient
LLCA Class Assignment	Class 2	112	41
	Class 1		19

K(172) = .376

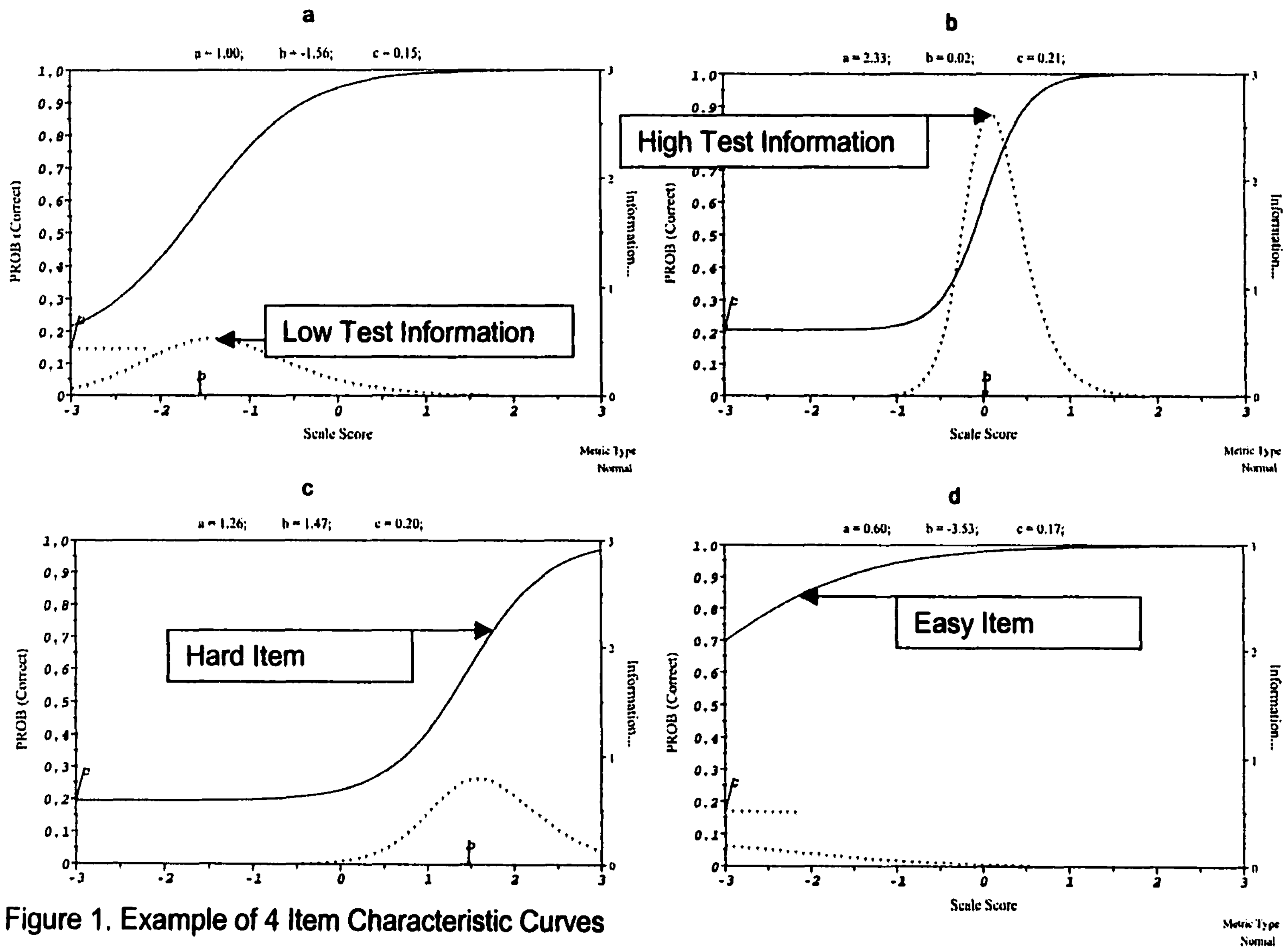
[a] N = 172, 20-item test

[b] Crosstabulation was the same using latent trait scores from the 2-, 3-, and 4-class model

Appendix B

Figures

Item Response Function and Item Information



Low Test Information

High Test Information

Hard Item

Easy Item

Figure 1. Example of 4 Item Characteristic Curves

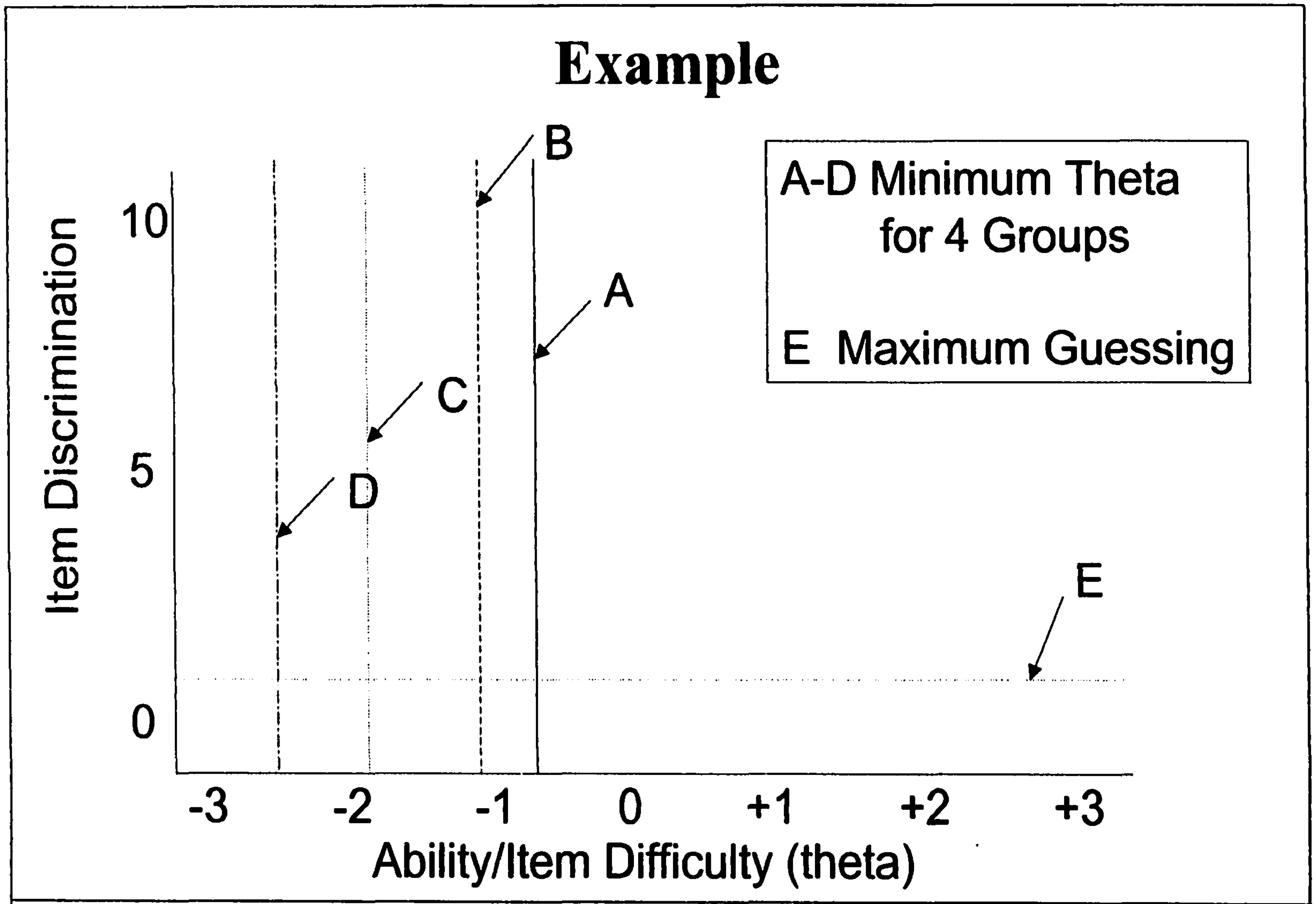


Figure 2. Example of Ability Boundaries

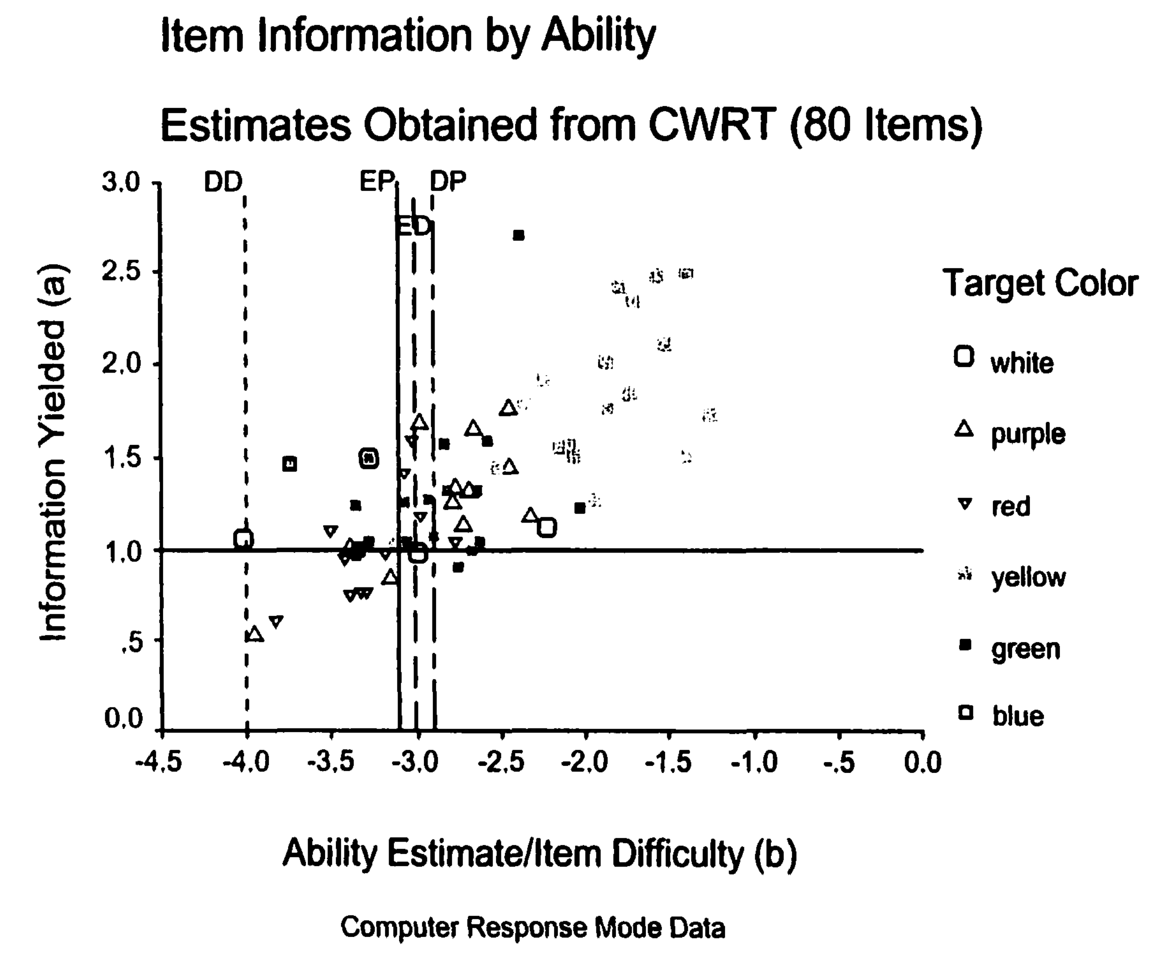


Figure 3. Item Content Color-Coded by Target Color

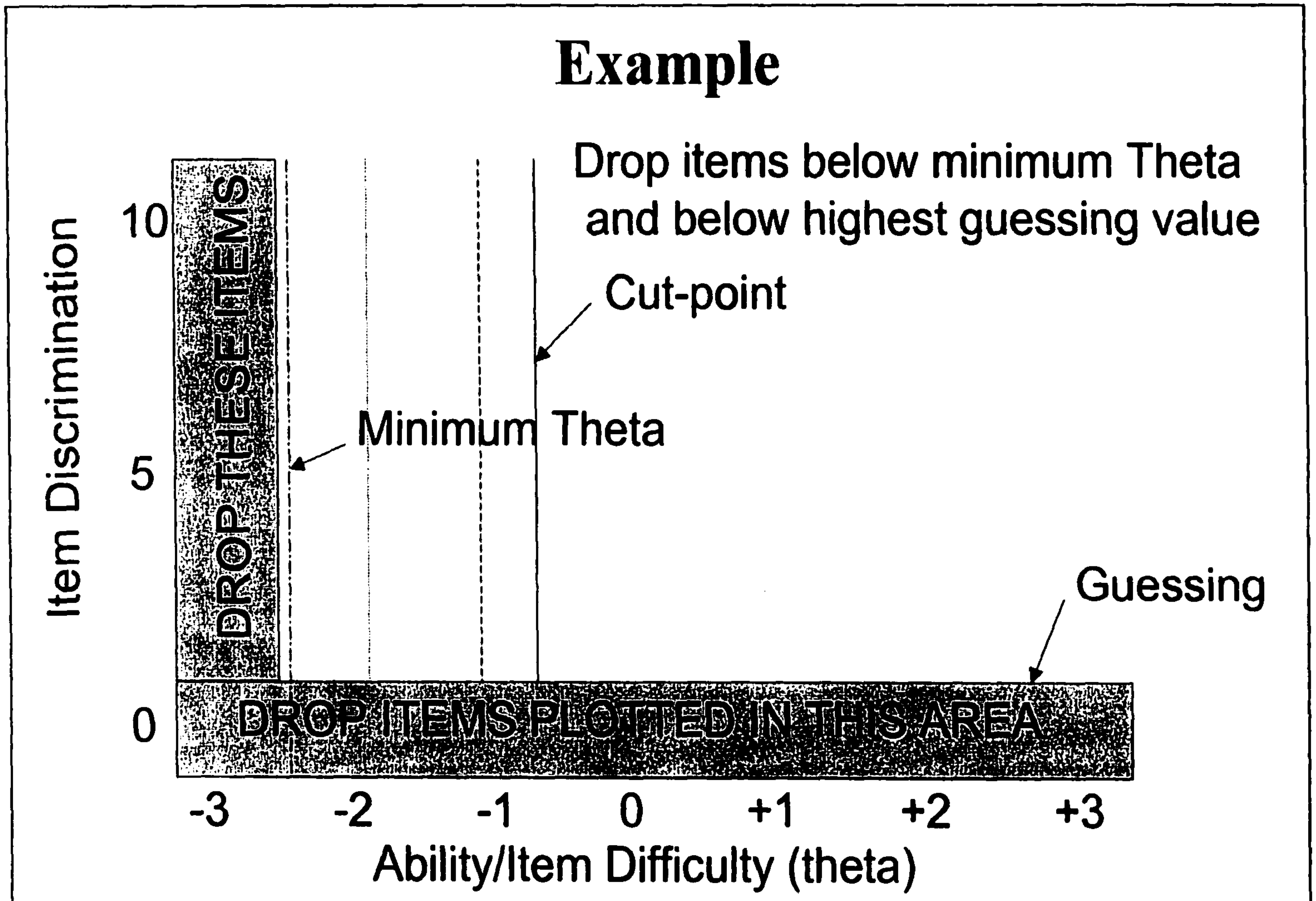

Example
Drop items below minimum Theta and below highest guessing value
Cut-point
Minimum Theta
Guessing
DROP THESE ITEMS
DROP ITEMS PLOTTED IN THIS AREA
Item Discrimination
10
5
0
-3
-2
-1
0
+1
+2
+3
Ability/Item Difficulty (theta)

Figure 4. Selecting Items for Omission

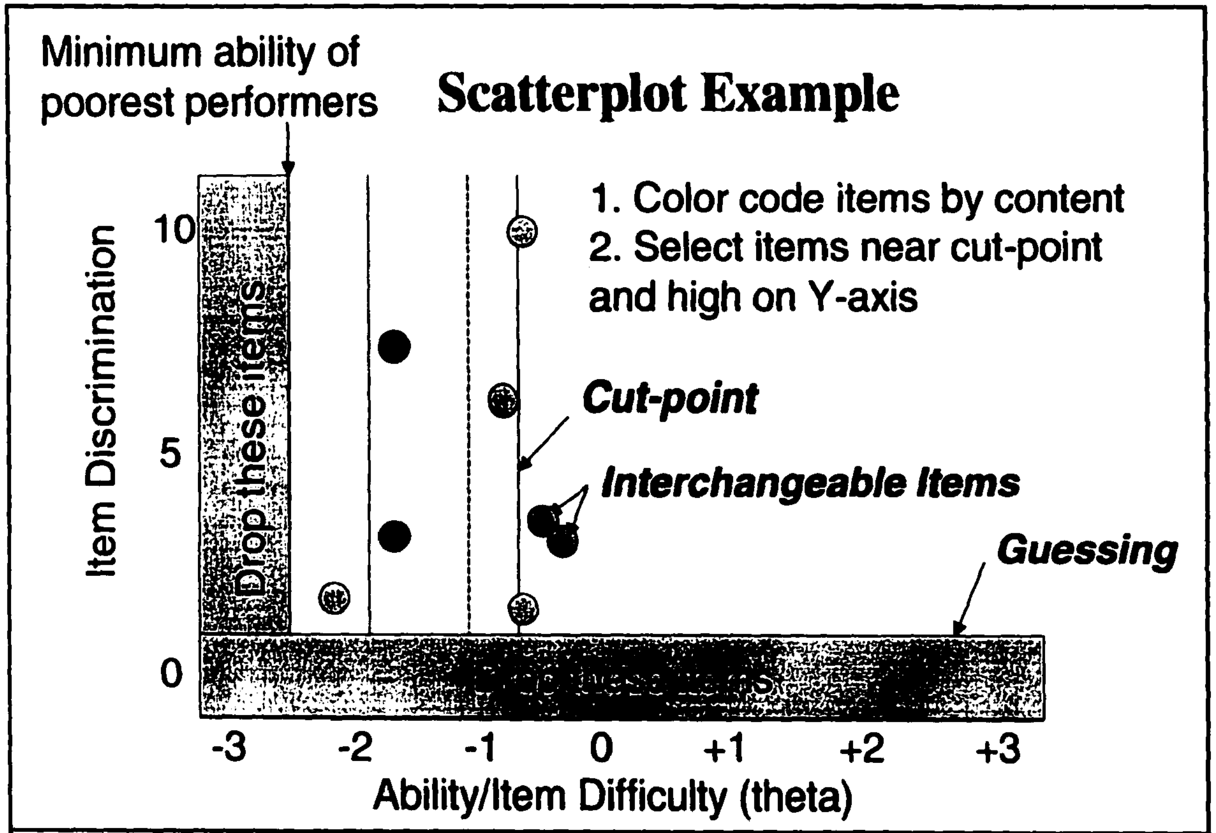


Figure 5. Selection Process

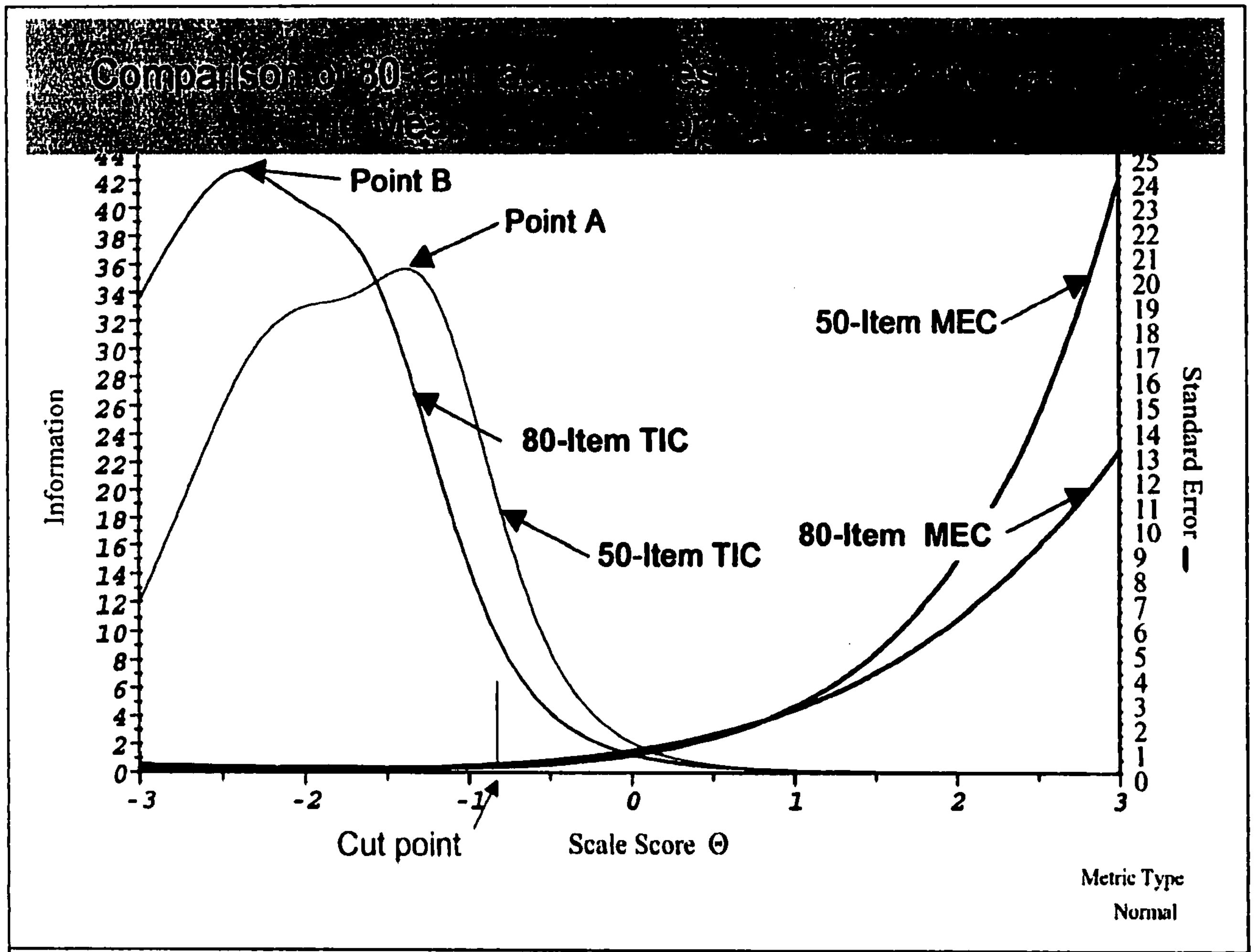

Comparison of 80-
Point B
Point A
50-Item MEC
80-Item TIC
50-Item TIC
80-Item MEC
Information
44
42
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Cut point
Scale Score Θ
Standard Error
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Metric Type
Normal

Figure 6.

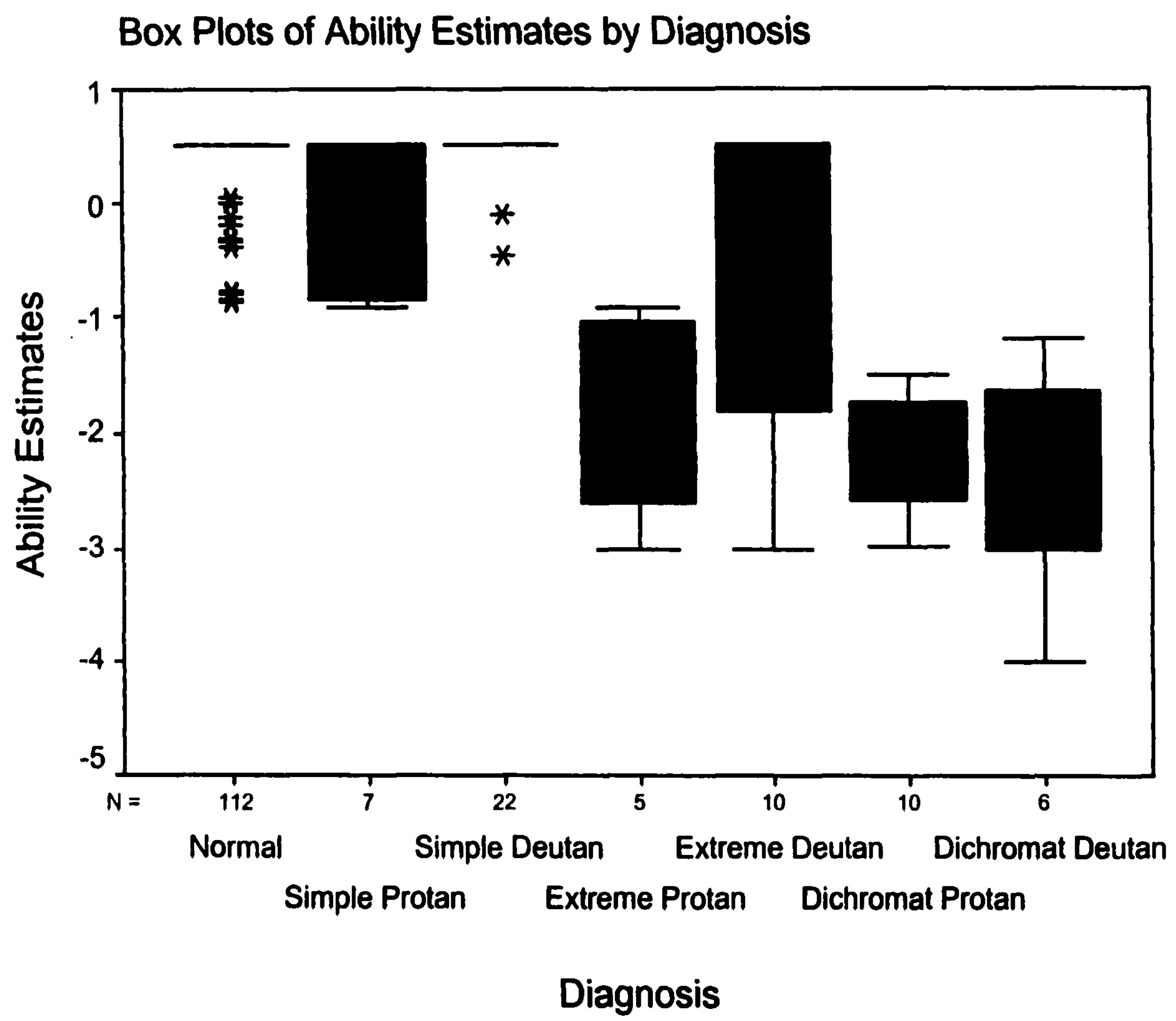


Figure 7. Boxplots of Ability Estimates by Diagnosis

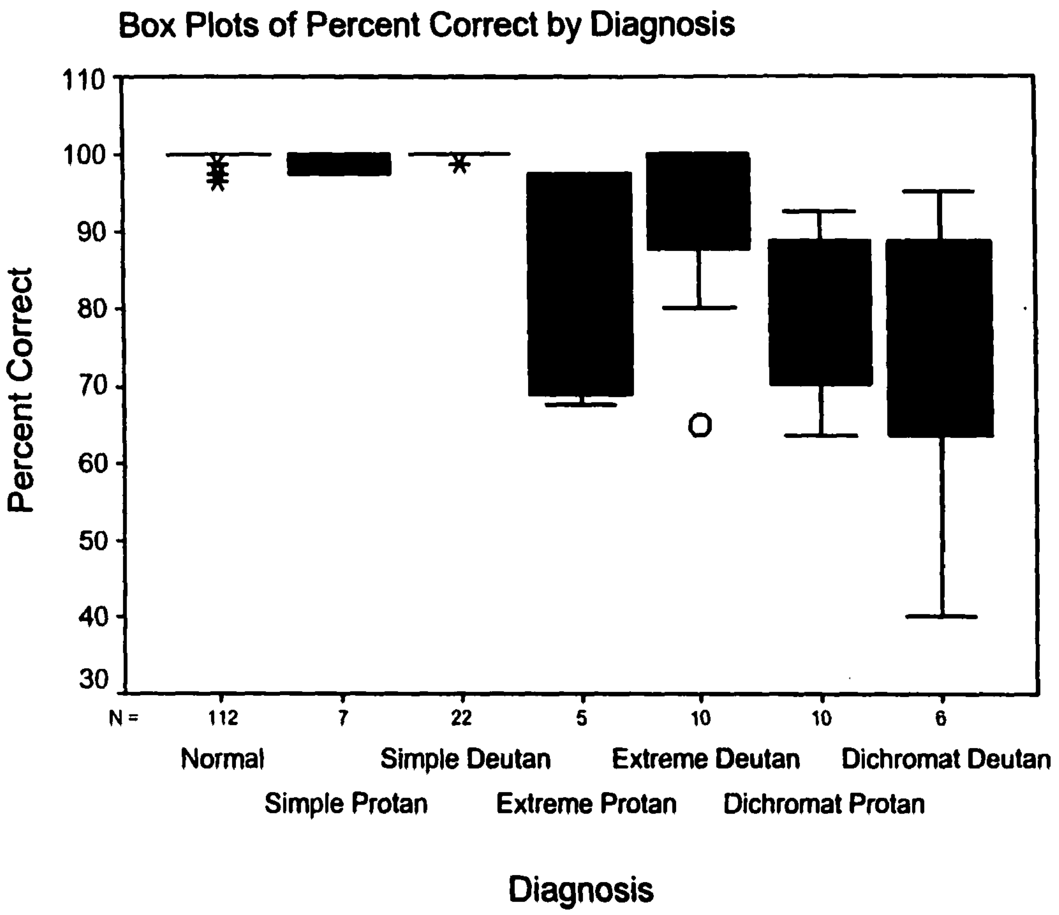


Figure 8. Boxplots of Percent Correct by Diagnosis

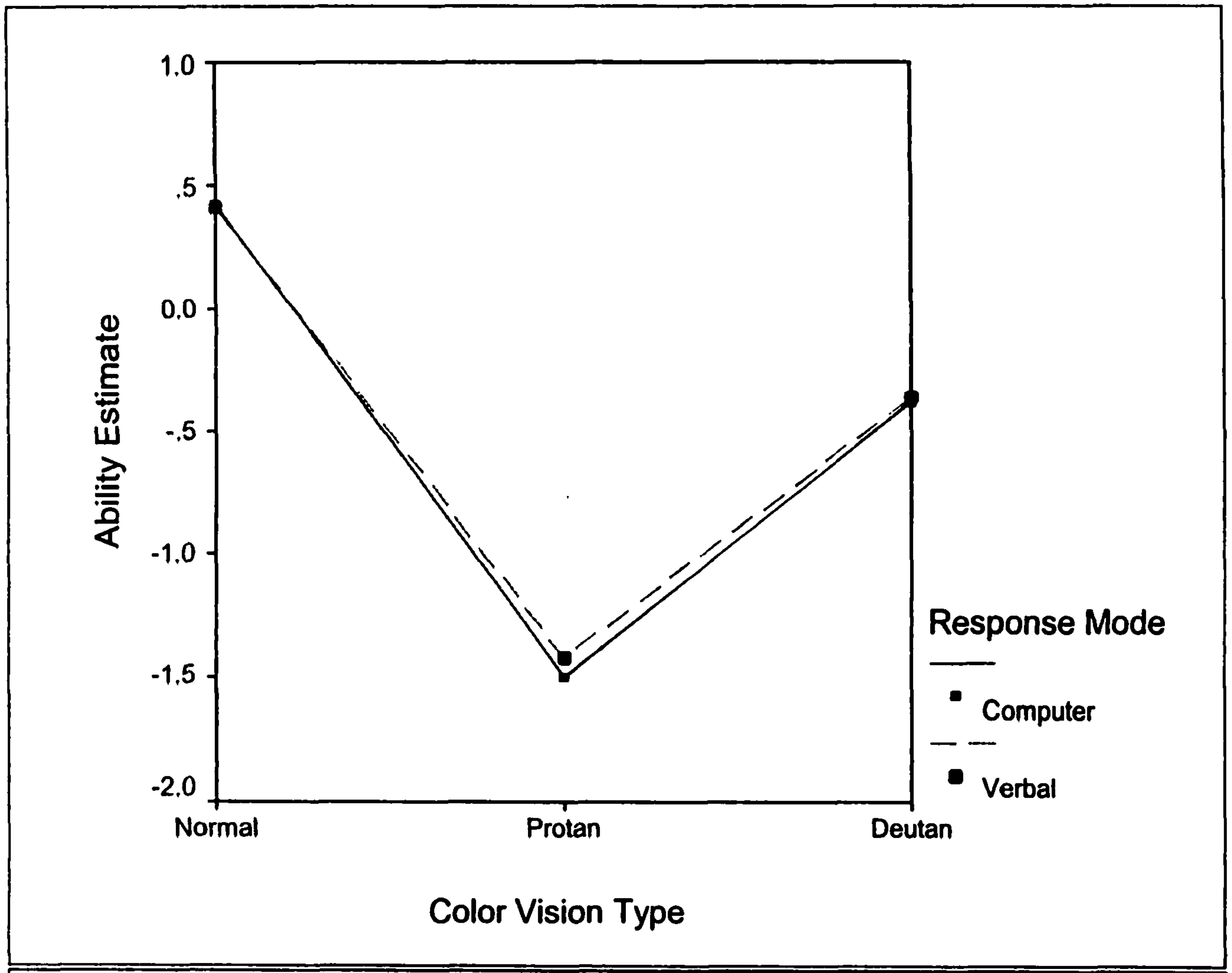


Figure 9. Ability Estimates for Computer and Verbal Response Modes as a Function of Type of Deficiency

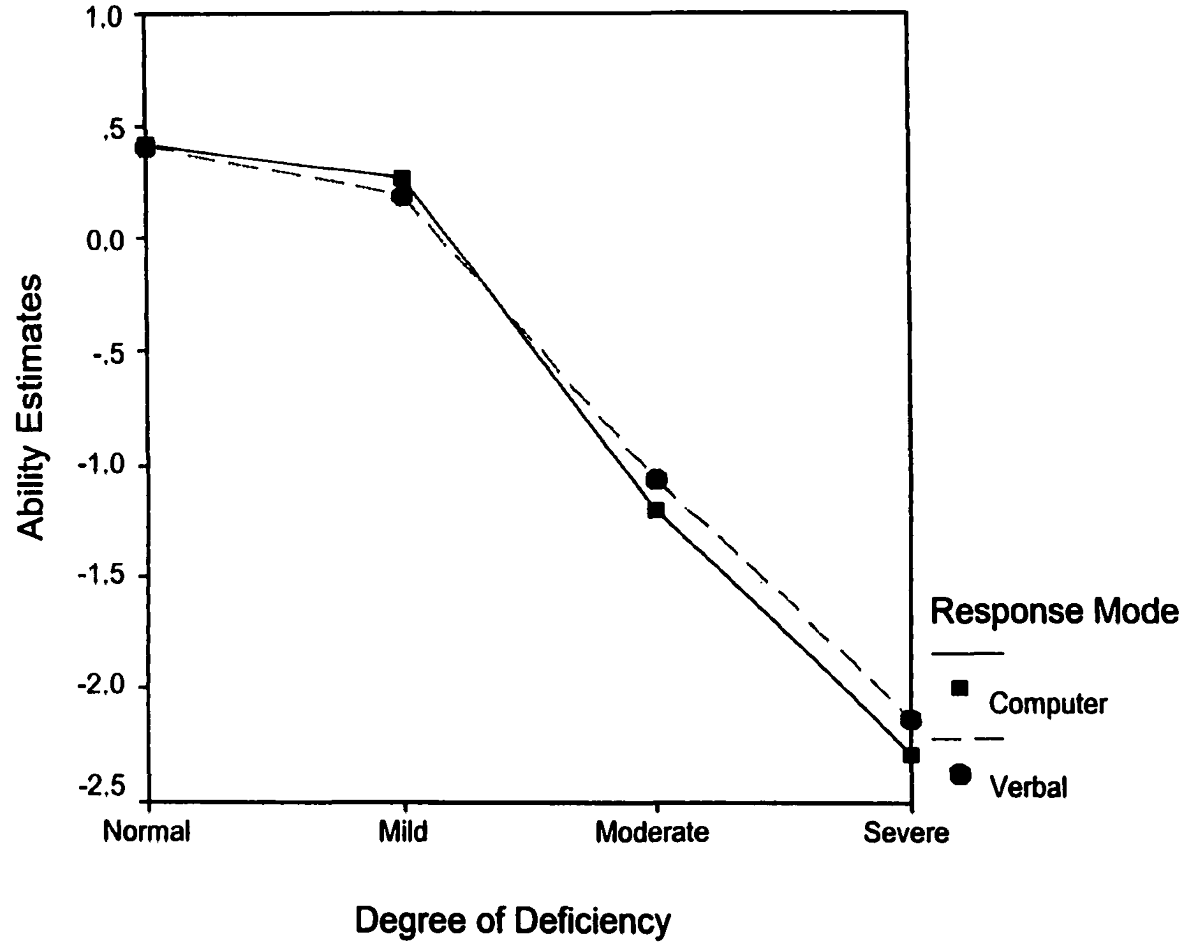


Figure 10. Ability Estimates for Computer and Verbal Response Modes as a Function of Degree of Deficiency

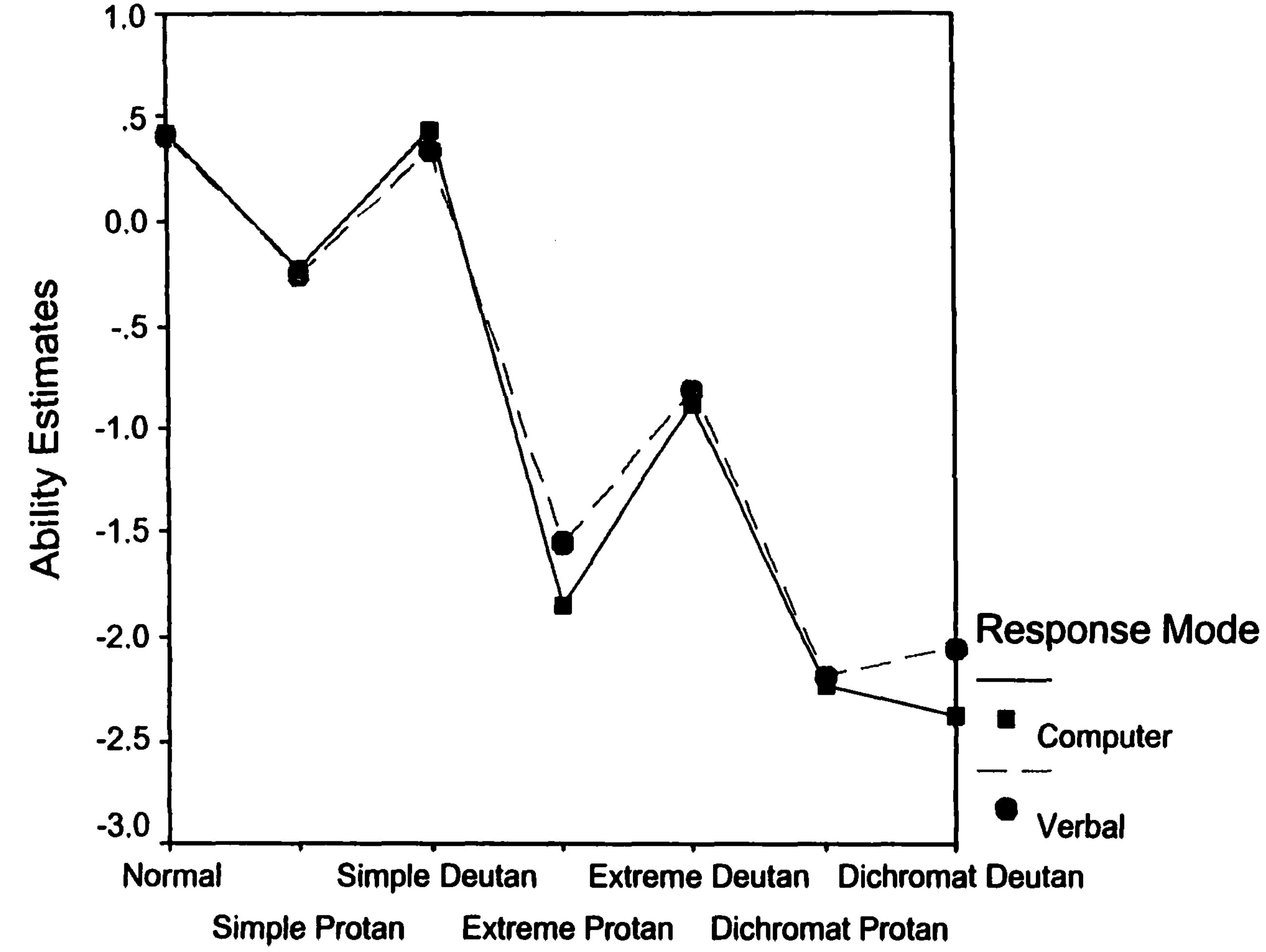
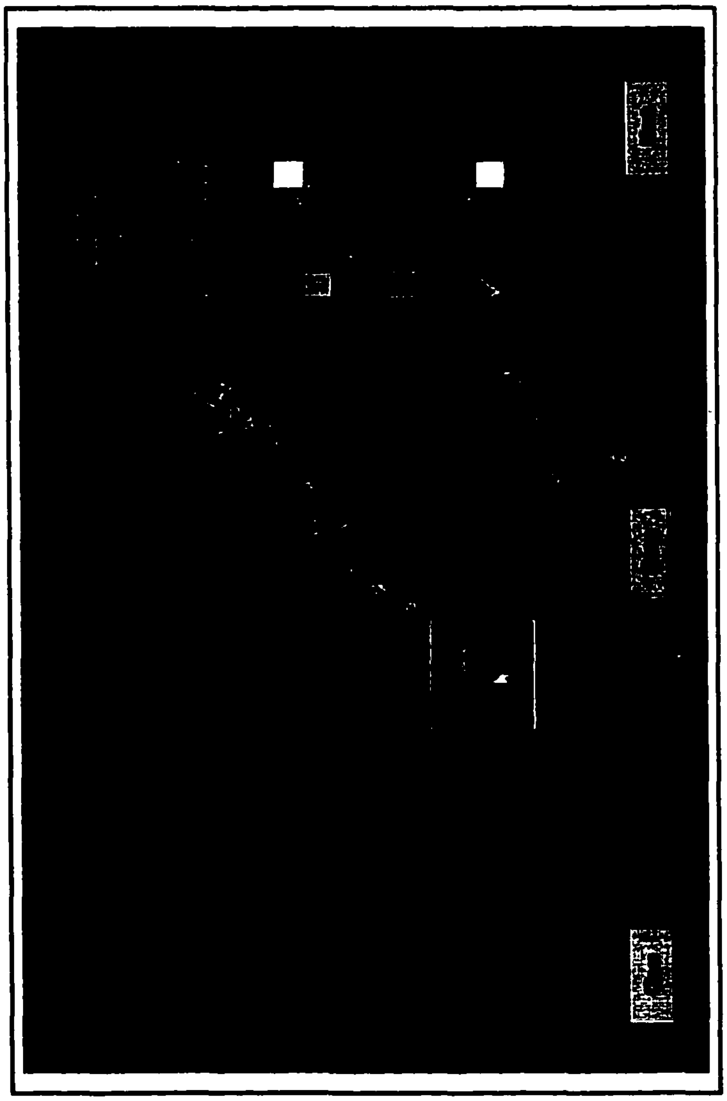


Figure 11. Ability Estimates for Computer and Verbal Response Modes as a Function of Diagnosis

Appendix C

Photograph



Photograph 1. Sample CWRT Trial