The Effect of *DIF* and Impact on Classical Test Statistics: Undetected *DIF* and Impact, and the Reliability and Interpretability of Scores from a Language Proficiency Test

Bruno D. Zumbo, Ph.D.
University of Northern British Columbia

Presented at the annual conference of the
National Council on Measurement in Education (NCME),
April, 2000, New Orleans, LA

Presented in the Invited Symposium "New Considerations of Validity with Second Language Tests" organized by Craig Deville and Micheline Chalhoub-Deville, University of Iowa.

Author information: Bruno D. Zumbo, Ph.D., is Full Professor of Psychology and Mathematics at the University of Northern British Columbia and Research Fellow at the School of Linguistics and Applied Language Studies at Carleton University. His primary interests are in psychometrics, statistical science, and the mathematical and philosophical foundations of measurement. Send correspondence to him at:

Department of Psychology,
University of Northern British Columbia,
Prince George, B.C.
V2N 4Z9 CANADA

e-mail: zumbob@unbc.ca
http://quarles.unbc.ca/psyc/zumbo/zumbo.html

Version: April 12, 2000

The Effect of *DIF* and Impact on Classical Test Statistics: Undetected *DIF* and Impact, and the Reliability and Interpretability of Scores from a Language Proficiency Test

Abstract: Based on the observation that classical test statistics are still in use in practice, this paper describes the results of a simulation study investigating the effect of differential item functioning (DIF) and impact on the classical test theory (CTT) statistics: coefficient alpha, alpha if item deleted, standard error of measurement, conditional standard error of measurement, and the corrected item-total correlation. The results indicate that DIF has no effect on the CTT statistics; however, impact modeled by an inequality of latent means and/or variances does inflate the reliability estimates and decrease the standard error statistics. The effect on the corrected item-total correlation is negligible. The results are discussed in the context of reliability and score interpretation.

Traditionally, the measurement properties of tests and measures in language testing have been summarized with omnibus statistics, such as reliability coefficients, item-total correlations, and standard error of measurement. These test statistics come together to characterize an approach to test development and evaluation often referred to as Classical Test Theory (CTT). Classical test theory has its intellectual history deeply rooted in the soil of correlation and group-based statistics (see, for example, Edgeworth, 1888, 1892; Spearman, 1904, 1907, 1913). What are currently considered classical test statistics are obviously far more highly evolved, both in practice and theory, than the original work of Edgeworth and Spearman, but today's CTT approaches are still rooted in the soil of group-based test- and score-level data and correlation.

These classical test statistics summarize the sample as whole and, in essence, average across levels of individual variation. Because these statistics summarize the sample as a whole, they ignore the fact that the measurement properties of the test may vary as function of variation within the sample. That is, the CTT methods provide a single global measure of test performance (e.g., coefficient alpha or conventional standard error of measurement) when, in fact, the test or measure may perform better at some score ranges than others. Moreover, because it is so strongly based on correlational measures it may provide an overly optimistic impression of the test's or measure's quality[1] at some score ranges. It is in this sense that these classical test statistics are considered omnibus summary measures. In contrast, approaches based on item response theory model the measurement properties as a function of the continuum of variation being assessed, such as language proficiency (see Hambleton, Swaminathan, & Rogers, 1991).

The growing interest in item response theory (IRT) by theoreticians and practitioners alike over the last 30 years has been nothing short of spectacular. This is evidenced in the number of sessions at measurement and testing conferences and the large proportion of publications in measurement and testing journals devoted to theoretical developments or applications of IRT. This evidence that IRT has tended to prevail in measurement journals and conferences would suggest that it is nearly exclusively used in measurement practice. Although it is true that IRT is being used quite a bit in large- to moderate-scale testing programs and projects, in some areas of the educational, social and behavioral sciences, the classical test statistics are widely used in the development and evaluation of tests and measures. This is clearly evident in areas of research and test development in which one is working with tests and

[1] In this sense, CTT is like much of ordinary least-squares regression which assumes, for example, a common variance along the regression line and hence reports one number for the standard error of estimate, irrespective of the value of the predictor(s).

measures of limited volume of production and distribution. For example, an overwhelming majority of tests and measures reviewed in source books such as the *Mental Measurements Yearbook* series produced by the Buros Institute of Mental Measurements or Robinson, Shaver, and Wrightsman's book *Measures of Personality and Social Psychological Attitudes* predominantly report CTT statistics. Likewise, in the field of language testing, some developers of small volume and research-oriented language tests use classical test statistics in the development and evaluation of their measures. The primary reason for using CTT in small volume testing programs and in research environments is the large sample sizes that are needed when one applies IRT[2].

I don't wish to be seen to slight IRT developments in any way with my previous comments. I wish only to highlight that CTT and score-level methods like generalizability theory (but more so CTT) are often being used in small volume and research-oriented tests and measures development and validation. As Brennan (1998) and others remind us, even though there have been many new approaches such as item response theory (IRT) developed, these classical test statistics are still used in practice -- in score interpretation and evaluating test score reliability. As Brennan writes in the journal *Educational Measurement: Issues and Practice*, "... classical test theory is alive and well, and it will continue to survive, I think, for both conceptual and practical reasons" (p.6).

It can be argued that the introduction of a "continuum of variation" into psychometric models that came with IRT has revolutionized measurement practice and has lead statistical thinking to an item-level analysis. In this light, one is, in essence, contrasting test or test score-level (i.e., CTT) versus item-level (i.e., IRT) approaches. I also believe that one of the positive benefits of taking into account the continuum of variation, and hence the focus on item-level psychometric models, has been the introduction of the notion of differential item functioning. As evidence for my beliefs, I point out that the last twenty years has seen a flurry of papers concerning definition, operationalization, detection, and measurement of DIF. It should be noted, however, that although the concept of DIF is a by-product of work in IRT and its concomitant injection of the continuum of variation into testing discourse, it is well known that there are also approaches to flagging DIF that do not depend on IRT.

Remarkably, even though classical and item response theories co-exist in meaurement theory and practice, there appears to be very little cross-talk between the two theories. The influential book by Lord and Novick (1968) demonstrates this nicely wherein it begins with one of the most sound and systematic discussions of classical models and then a shift is made later in the book to item response theory with no discussion of how one affects the other. Furthermore, most introductions to IRT motivate the need for IRT by briefly describing the limitations of the classical model but then never return to the classical model again. As Brennan (1998) states, it is as if there is a tendency to think of the newer models as a replacement for classical theory when they really should be seen as extensions. Likewise, CTT specialists generally do not work on IRT problems and hence do not concern themselves with issues such as DIF. Given that this paper is a cross-over paper between IRT and CTT, Appendix A provides a brief introduction to DIF to help readers who are not familiar with DIF to understand the design and results of this study.

---

[2] The exception to this sample size concern is the program TESTGRAF98 (Ramsay, 1999) which eliminates the large sample size requirements of standard IRT methods.

Based on the notion that classical test statistics are still in use, this paper is the first of a research program devoted to studying how developments and ideas in item response theory affect classical test theory -- particularly in the context of language proficiency testing. In essence, I am looking at the influence of ideas in IRT on classical test statistics. Specifically, this paper will examine how the classical test statistics are affected by the presence of undetected DIF -- i.e., DIF exists but the test developer or user does not know that it exists because in the CTT framework they are not thinking in terms of DIF issues. The results will be discussed in the context of test score interpretation and reliability. In the end, this speaks to construct validity evidence by examining the effect of undetected DIF on classical test statistics.

The paper will report on a simulation study that examines the effect of varying levels of DIF and impact on classical test statistics. A test will be simulated to reflect the measurement of English proficiency of individuals whose native language is not English. To this end, I will be using item parameter estimates for ***TOEFL*** to simulate the data in the various conditions. More specifically, I will be using the item parameter estimates for the *Structure and Written Expression* section of the paper-and-pencil version of the ***TOEFL***. The simulated test will then allow me to examine the effect of undetected DIF on classical test statistics applied to a test of knowledge of structural and grammatical elements of standard written English.

Methodology

Item Parameters Used in the Simulation

The Test of English as a Foreign Language (***TOEFL***) is made up of several sections and contains a mixture of item types. The ***TOEFL*** traditionally has been comprised of three separately timed sections: listening comprehension, structure and written expression, and vocabulary and reading comprehension. In this simulation study, I will be using real IRT item parameter statistics from the structure and written expression section of the ***TOEFL***. My reasons for using this information in my simulation are that this section of the ***TOEFL***:

- is composed of independent binary scored items.
- Is similar to several language tests currently being used in research and small-scale language testing and it consists of sentences that test one's knowledge of important structural and grammatical elements of standard written English.

In its essence, however, my reasons for using these item parameters are that the item format, difficulty, and discrimination ranges reflect realistic test properties in language testing. I would like to thank ETS, and Dan Eignor in particular, for providing the item parameter statistics for the 38 item version of the structure and written expression section of the ***TOEFL***. Appendix B lists the item parameters for the 38 items.

Simulation Design

Examinee response data were simulated using MIRTGEN 1.0 (Luecht, 1996) using a three parameter logistic item response model, under a variety of test conditions. These test conditions included ability distribution differences that were reflected in both the mean and variance of the latent distributions, and in the levels of DIF which ranged from no DIF to a large magnitude of DIF. Although MIRTGEN allows for multivariate (i.e., multidimensional) IRT, I studied the more commonly found univariate (i.e., unidimensional) IRT model.

Given that this is the first study of a much larger research program, several other variables that are discussed in the DIF literature were not studied in this simulation. That is, the following variables were held to a fixed level (i.e., not manipulated) in the simulation study.

- Sample size: This simulation does not focus on the re-sampling estimate of the standard error of the CTT statistics. Instead, I am focusing on the population values of the CTT statistics and hence I am estimating these statistics with sample sizes of 10,000 examinees. This, of course, implies that there is not a need for more than one replication per cell because the standard error of these CTT statistics over repeated samples would be quite small with such a large sample size.
- Sample size per group: I have limited my study to equal sample sizes of 5,000 in each group.
- Varying test length: I am studying a fixed test length of 38 items. I could have generated multiples of the 38 items to increase test length, but I wanted to keep the number of items in the test limited to those that reflect the standardized tests in research or small-scale testing programs.
- Percentage of items containing DIF: I have limited the simulation to one item with DIF. This is the boundary experimental condition to investigate whether even one DIF item will have an effect on the CTT statistics.
- Type of DIF, uniform or non-uniform: I will focus on uniform DIF only. This reflects a situation in which DIF may be attributable to differences in item difficulty only.
- As in most measurement studies, the latent distribution is considered a Normal (i.e, Gaussian) distribution.

The following design variables were manipulated in the simulation.

- Item impact in the form of differing latent means (see Appendix A): Impact is defined as the difference in latent distributions (i.e, the distribution of the unobserved continuum of variation). As is sometimes done in DIF studies, I simulated conditions under which the two groups have equal means (i.e., no impact) and conditions under which Group One has a mean of zero and Group Two has a mean of 1 (i.e., impact).
- Item impact in the form of differing latent variances: As Zumbo and Coulombe (1997) state, observational studies, of which DIF studies are an example, may result in unequal variances for the two groups. That is, they state "… there are two situations in which one cannot necessarily assume equality of variances: (a) the groups of experimental units are formed by domain differences such as age groups, gender, or educational level, and/or (b) when the experimental units (knowingly or unknowingly to the experimenter) differ on some important, possibly unmeasured, variable" (p. 147). In the previous quotation, "experimental units" are examinees in a DIF study. Furthermore, although the above two conditions describe DIF studies, this factor is seldom, if ever, studied in DIF research. For the simulation, Group One always has a variance of 1.0 whereas Group Two was manipulated to have either a variance of 1.0 or a variance of 4.0.
- Magnitude of DIF: Three levels of DIF magnitude were studied,
    1. negligible DIF, a A-level DIF defined by ETS (Zwick & Ercikan, 1989),

2. a difference in difficulty -- i.e., b-parameter -- of 0.50 which corresponds to a moderate (B-level) amount of uniform DIF (as defined by ETS), and
3. a difference in difficulty -- i.e., b-parameter -- of 1.0 which corresponds to a large (C-level) amount of uniform DIF (as defined by ETS).

The resulting research design was a 2 x 2 x 3 design with one replicate per cell. The design factors represent latent mean differences, latent variance differences, and magnitude of DIF, as described above. A graphical depiction of the design can be seen in Figure 1.

Figure 1. Study Design

	Equal population means for the two latent distributions $Mean_1=0$; $mean_2=0$	Unequal population means for the two latent distributions $mean_1=0$; $mean_2=1$
Equal population variances for the two latent distributions $var_1=1$; $var_2=1$	Group 1: $N(0,1)$, Group 2: $N(0,1)$ Magnitude of DIF None Moderate Large	Group 1: $N(0,1)$, Group 2: $N(1,1)$ Magnitude of DIF None Moderate Large
Unequal population variances for the two latent distributions $var_1=1$; $var_2=4$	Group 1: $N(0,1)$, Group 2: $N(0,4)$ Magnitude of DIF None Moderate Large	Group 1: $N(0,1)$, Group 2: $N(1,4)$ Magnitude of DIF None Moderate Large

Note: The notation Group #: $N(\#, \#)$ denotes a Normal distribution with a given mean and variance. Therefore, the upper left-hand cell describes a simulated test with equal means (zero) and variances (one) for the two latent distributions, and hence no impact.

How was DIF modeled?

Recall that the DIF being studied herein is only attributable to differences in the item difficulty (b-parameters). Item #10 which had a b-parameter of -0.2517 was chosen to serve as the target item in which DIF will be modeled. Table 1 lists the corresponding b-values for the three levels of DIF. Note that Group 2, will reflect any impact effect on means and variances of the latent distribution and DIF – i.e., difference in b-parameters whereas Group 1 always remains the same.

Table 1. The b-parameters the three levels of DIF.

Magnitude of DIF	**Difficulty (b-parameters) for Group 1**	**Difficulty (b-parameters) for Group 2**
Level A, No DIF	-0.2517	-0.2517
Level B, Moderate DIF	-0.2517	0.2485
Level C, Large DIF	-0.2517	0.7485

What were the dependent variables in the study?

For each of the 12 cells in the study design, the test reliability (coefficient alpha) was computed, as well as the corrected item-total correlation, and alpha if item is deleted for the study item (#10). In addition, the traditional CTT standard error of measurement (SEM) was computed using the well-known formula

$$SEM = sd\ \sqrt{1 - reliab},$$

where ***sd*** denotes the standard deviation for the observed total scores (the simple sum for the 38 items) for the entire examinee group (in our case 10,000 simulated examinees) and ***reliab*** denotes the reliability estimate, in this case coefficient alpha.

Finally, because language tests are often used for selection purposes, the conditional standard error of measurement was also computed. It should be noted that the conditional standard error of measurement is a measure that takes into account the latent continuum of variation and hence is not really a part of what is commonly called classical test theory. However, because the Standards for Educational and Psychological Testing have suggested reporting the conditional standard error of measurement, I decided to include this conditional measure as well. Therefore, to investigate the impact of DIF on the conditional standard error of measurement, the software TESTGRAF98 (Ramsay, 1999) was used to compute the conditional standard error of measurement. If we assume that the cut-score for this test is 20 (out of a possible 38) points, the conditional standard error of measurement was computed for the expected score of 20. Please note that this cut score was selected because under the no-DIF and no impact condition, the study item (item #10) measures optimally at an expected score of 20. That is, the item, in a technical sense, provides maximal statistical information or precision for the estimate at an expected score of 20. Therefore, I am studying the effect of undetected DIF on the conditional standard error at the point of the continuum of variation where the items works best and (by design and not by coincidence) at the cut-score for the test.

The conditional standard error of measurement (CSEM) was computed using the formula

$$CSEM = 1/\sqrt{I(20)},$$

where $I(20)$ denotes the statistical information at the expected score of 20 (out of 38). The information was computed by the TESTGRAF98 program.

Each of these statistics, the reliability estimate, the group-based standard error of measurement, and the conditional standard error of measurement, the corrected item-total correlations, and the alpha if item is deleted were computed based on the 10,000 respondents for each of the 12 cells of the study design.

Results

Figure 2 lists the results of the five outcome variables for the 2x2x3 design described above. Treating each outcome measure separately, each cell of the 2x2x3 design has one observation. As Kirk (1982, p. 399) states, in the case of the analysis of completely randomized factorial designs with n=1 per cell, the highest-order interaction, instead of the within-cell error term, is used as an estimate of the experimental error.

Figure 3, therefore, lists the results of the 2x2x3 ANOVAs with the three-way interaction term as the experimental error for each outcome variable. For the three correlation-based outcome variables (alpha, alpha if item deleted, and corrected item-total correlation) a Fisher's r-to-z transformation was performed on the values before the ANOVA was conducted. Upon inspection of Figure 3, it becomes apparent that the DIF factor had no statistically significant effect on its own or as an interaction with the two factors for any of the outcome CTT statistics. The only statistically significant effect on CTT statistics resulted from the impact factors of mean differences, variance differences or the interaction between these two impact variables.

To help interpret the results in Figure 3, Figure 4 reports the cell means in terms of deviation from the baseline values (see the bolded column in Figure 2). That is, given that the DIF factor was statistically non-significant, Figure 4 reports the cell and marginal means for the two impact factors collapsing (i.e., averaging) over the DIF factor (i.e., for each CTT test statistic, each cell of the 2x2 design is based on 3 observations). Therefore, the values in Figure 4 represent the magnitude of the effect in terms of how much the impact factors alter the CTT statistics from their baseline values. Figure 4 reports the cell means for the standard error of measurement and conditional standard error of measurement. Furthermore, to aid in interpretation of the correlation-based measures (i.e., reliabilities and corrected item-total correlations), I transformed the difference in z-scores back to the correlation value.

Figure 2. Results

	Equal population means for the two latent distributions $mean_1=0$; $mean_2=0$	Unequal population means for the two latent distributions $Mean_1=0$; $mean_2=1$
Equal population variances for the two latent distributions $var_1=1$; $var_2=1$	Group 1: *N*(0,1), Group 2: *N*(0,1) Magnitude of DIF **None** Moderate Large SEM **2.588** 2.588 2.588 CSEM **1.186** 1.189 1.180 Reliab **.876** .876 .875 Alpha-del **.871** .873 .872 C-ITC **.462** .392 .400	Group 1: *N*(0,1), Group 2: *N*(1,1) Magnitude of DIF None Moderate Large SEM 2.409 2.408 2.408 CSEM 1.217 1.202 1.197 Reliab .900 .899 .899 Alpha-del .896 .894 .894 C-ITC .529 .551 .529
Unequal population variances for the two latent distributions $var_1=1$; $var_2=4$	Group 1: *N*(0,1), Group 2: *N*(0,4) Magnitude of DIF None Moderate Large SEM 2.509 2.509 2.508 CSEM 1.266 1.265 1.266 Reliab .919 .919 .919 Alpha-del .916 .916 .916 C-ITC .539 .571 .538	Group 1: *N*(0,1), Group 2: *N*(1,4) Magnitude of DIF None Moderate Large SEM 2.399 2.410 2.400 CSEM 1.046 1.048 1.056 Reliab .925 .923 .926 Alpha-del .922 .920 .923 C-ITC .564 .569 .569

Note: The bolded column is the base-line condition with no DIF, or impact of either mean or variance. The notation Group #: *N*(#, #) denotes a Normal distribution with a given mean and variance. Therefore the upper left-hand cell describes a simulated test with equal means (zero) and variances (one) for the two latent distributions. Also Reliab = coefficient alpha, SEM = standard error of measurement, CSEM = conditional standard error of measurement at the expected score of 20, C-ITC = corrected item-total correlation, and Alpha-del = alpha if item deleted.

Figure 3. ANOVA results for the outcome variables

Statistical Effect in the ANOVA	**OutcomeVariables**				
	Alpha	Alpha if item deleted	Corrected item-total correlation	Standard error of measurement	Conditional standard error of measurement at a score of 20
1. Impact modeled by a difference in means of the latent variable	✓	✓	×	✓	✓
2. Impact modeled by a difference in variances of the latent variable	✓	✓	✓	✓	✓
3. DIF modeled by a difference in item difficulties	×	×	×	×	×
Interaction of factors 1 by 2	✓	✓	×	✓	✓
Interaction of factors 1 by 3	×	×	×	×	×
Interaction of factors 2 by 3	×	×	×	×	×

Note: The symbol ✓ denotes a statistically significant effect at a .05 significance level, whereas a × denotes that the effect was not statistically significant.

Figure 4. The cell means of the deviation scores (from the base-line condition) by the two impact factors (means and variances of the latent distribution).

SEM	$Mean_1=0$; $Mean_2=0$	$Mean_1=0$; $Mean_2=1$	
$Var_1=1$; $Var_2=1$	-.00015	-.1796	-.09
$Var_1=1$; $Var_2=4$	-.079	-.1849	-.1319
	-.04	-.1823	

CSEM	$Mean_1=0$; $Mean_2=0$	$Mean_1=0$; $Mean_2=1$	
$Var_1=1$; $Var_2=1$	-.00075	-.0197	-.0095
$Var_1=1$; $Var_2=4$	-.0802	-.1355	-.0276
	-.0397	-.0579	

Reliability	$Mean_1=0$; $Mean_2=0$	$Mean_1=0$; $Mean_2=1$	
$Var_1=1$; $Var_2=1$	.0001	.0431	.0241
$Var_1=1$; $Var_2=4$	.0976	.1140	.1058
	.0490	.0813	

Alpha if item deleted	$Mean_1=0$; $Mean_2=0$	$Mean_1=0$; $Mean_2=1$	
$Var_1=1$; $Var_2=1$	.0019	.0468	.0243
$Var_1=1$; $Var_2=4$	.0908	.1137	.1059
	.0501	.0803	

Corrected item-total correlation	$Mean_1=0$; $Mean_2=0$	$Mean_1=0$; $Mean_2=1$	
$Var_1=1$; $Var_2=1$	-.0230	.0431	.0098
$Var_1=1$; $Var_2=4$	.0512	.0623	.0567
	.0139	.0527	

I will now discuss each of the five CTT outcome measures.

Standard error of measurement: In Figure 4, the main effect of the latent variable means factor can be clearly seen by contrasting the marginal effects of -.04 when there is no difference between the latent means and -.1823 when the latent means differ. Therefore, impact, reflected in a difference between the means of the latent distribution, results in a decrease of the standard error of measurement. This effect is also slightly larger for the equal variance condition compared to the unequal variance condition, which also explains the main effect of variance inequality.

Conditional standard error of measurement at the cut-score: From Figure 4 it is apparent that the only condition where there is a meaningful effect of impact for practical purposes is the condition wherein the impact is seen in a difference in latent means and latent variances between the two groups. All of the other effects were, for practical purposes of testing, negligible. That is, the next largest effect was for equal latent means but unequal latent variances with the conditional standard error was reduced by .0802, on average, where we can see in Figure2 that the conditional standard errors were in the range of 1.0 to 1.5.

Reliability: Again from a practical point of view, the main effect of latent variance inequality resulted in an inflation of the reliability estimate (coefficient alpha), as did the main effect of latent mean differences. The effect, however, was more pronounced for the latent variance factor in part because the condition of unequal latent means and unequal latent variances had the most dramatic effect by inflating the alpha by .1140, on average.

Alpha if item #10 (the study item) is deleted: This result is the same as the result for the coefficient alpha reliability described above.

Corrected item-total correlation for item #10 (the study item): Given that the original magnitudes of the corrected item-total correlations were in the range of .40 to .57, the main effect of latent variance inequality was considered, for practical purposes, to be negligible.

Discussion

I will discuss the results in the context of a realistic testing setting. Imagine you are developing a 38-item language proficiency test with a cut-score for language proficiency for academic purposes of 20. The items are independent and have a binary (right/wrong) scoring format. Your testing population includes European educated and Asian educated students. You are concerned that the European students, because they are not used to multiple choice testing situations and examination formats, may be disadvantaged. You begin to learn about IRT specialists' concern for DIF and impact. You, of course, do not have the resources or sample size to consider using IRT in your test development practice; instead you rely on the commonly used CTT approach. However, as you learn about DIF and impact you ask yourself the question: what effect might DIF or impact have on the CTT test statistics you use in your everyday test development?

Although the results herein are preliminary, you might be able to say:

1. If there is no impact, the effect of having different item difficulties (i.e., DIF) is negligible for any of the classical test statistics (i.e., coefficient alpha, alpha if item is deleted, standard error of measurement, and the corrected item-total correlation). Furthermore, the conditional

standard error of measurement at the cut-score of 20 is also not affected. This means that if there is no impact the reliability and interpretability of the test scores are not affected by this sort of DIF.

2. Another interesting consequence of the finding that DIF has no effect on the CTT statistics is that the commonly heard comment that DIF affects the test reliability seems to not hold up to empirical testing, at least not in the conditions studied herein.

3. What does seem to affect the CTT statistics, however, is impact. That is, if the European and Asian students have a difference in probability of responding correctly to the language proficiency item because there are true differences between the groups in the underlying language proficiency being measured by the item, then the CTT statistics are affected. In short, the coefficient alpha and the alpha if item deleted will be slightly inflated, and (as expected) the standard error of measurement and conditional standard error of measurement are decreased. The effect on the corrected item-total correlations is negligible.

4. The results regarding the main effects of impact are explainable from basic theory of reliability. That is, as is well known, let us define reliability as

$$reliability = \frac{\text{var}(\theta)}{\text{var}(\theta) + \text{var}(\varepsilon)}, \qquad (1)$$

where, for example, in the case of our simulation var(θ) and var(ε) are computed for the 10,000 scores and denote the variance of the ability and error, respectively. As is well known, if the error variance is fixed, an increase in the variance of the true score, in our case θ, will result in increased reliability. We can see from equation (1) that, assuming var(ε) is held constant, if one increases var(θ) (relative to the no impact condition) as a result of the inequality of variances for the two sub-groups (each of 5,000 scores), the reliability will increase. Likewise, the same effect happens when there is an inequality in latent means. Note that this reminds us that reliability is inflated by impact only in the relative sense as compared to the no impact condition -- that is, the statement that a quantity is "inflated" is a comparative statement. In its essence, impact is increasing var(θ) and hence inflating reliability.

In closing, a unique component of this study that needs highlighting is that impact can be studied as an inequality of latent means and/or latent variances. The co-occurrence of inequality of latent means and latent variances, however, can have a particularly strong effect on the CTT statistics. On-going research is expanding the facets of the simulation design to include the variables described in the methodology section and to investigate the standard error across replications of the simulation for smaller sample sizes.

References

Brennan, R. L. (1998). Misconceptions at the Intersection of Measurement Theory and Practice. *Educational Measurement: Issues and Practice, 17*, 5-9, 30.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599-635.

Edgeworth, F. Y. (1892). Correlated averages. *Philosophical Magazine, 34,* 190-204.

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

Kirk, R. E. (1982). *Experimental Design (2nd Ed.)*. Belmont Calif: Wadsworth INC.

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

Luecht, R. (1996) MIRTGEN 1.0 [Computer Software]. Author.

Ramsay, J. O. (1999). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data*. Author: McGill University.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161-169.

Spearman, C. (1913). Correlatins of sums and differences. *British Journal of Psychology, 5*, 417-426.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51,* 139-150.

Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57*, 963-969.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.

Appendix A

A Brief Introduction to *DIF*

The term *differential item functioning* (DIF) is often interpreted as an indication of whether or not the same psychological constructs are measured across different groups. For example, if an item does not measure the same skills or subskills in different populations, it is said to function differentially or to display item bias. The following material is based on my DIF Handbook (Zumbo, 1999).

Some Definitions:

Item impact: Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item because there are true differences between the groups in the underlying ability being measured by the item.

DIF: DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying ability* that the item is intended to measure.

Item bias: Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose. DIF is required, but not sufficient, for item bias.

Item impact and item bias differ in terms of whether group differences are based on relevant or irrelevant characteristics (respectively) of the test. DIF requires that members of the two groups be matched on the relevant underlying ability before determining whether members of the two groups differ in their probability for success.

DIF is a necessary, but not sufficient, condition for item bias. Thus, if DIF is *not* apparent for an item, then no item bias is present. However, if DIF is apparent, then its presence is *not* sufficient to declare item bias; rather, one would have to apply follow-up item bias analyses (e.g., content analysis, empirical evaluation) to determine the presence of item bias.

As Clauser and Mazor (1998) remind us, it is important to distinguish both item bias and DIF from inappropriate item content or framing that is potentially offensive. What is important to keep in mind is that if an item is offensive to everyone, it is not going to be detected as biased -- by the simple fact that lack of bias means that no one group is affected but rather both (or all) groups are equally affected.

Fundamentals of IRT: Statistical Methods For Item Analysis

In this section I will discuss further the continuum of variation of a latent variable. The continuum of variation is an essential part of modern test theory and it not only helps define DIF but it also helps us interpret DIF results.

A latent variable represents a quantity. That is, respondents have an amount of this latent variable. For example, an amount of intelligence, verbal ability, spatial ability, sociability, etc., and that it is a continuum along which individuals vary. That is, different people may have a different amount of the latent variable we are trying to tap with our items. The term continuum

of variation is far less politically laden than the term “latent trait.” This concept of a continuum of variation is fundamental to modern test theory.

Modern test theory, in general, according to Hambleton, Swaminathan, and Rogers (1991), is based on two postulates:

1. that the performance of an examinee on a test item can be explained or predicted from a set of factors traditionally called "traits, latent traits, or abilities" -- which we refer to as a continuum of variation; and
2. that the relationship between examinees' item performance and the continuum of variation underlying performance on that item can be described as an item characteristic curve (ICC) or item response function (IRF).

A parametric ICC is the monotonically increasing function to which items are fit in most item response models. Historically, the function was the normal ogive function but was later replaced with the more tractable logistic function. Parametric ICCs vary in terms of their position on the X-axis, their slope, and their intercept with the Y-axis. The X-axis is the latent variable and the Y-axis is the probability of getting the item correct (or endorsing the item).

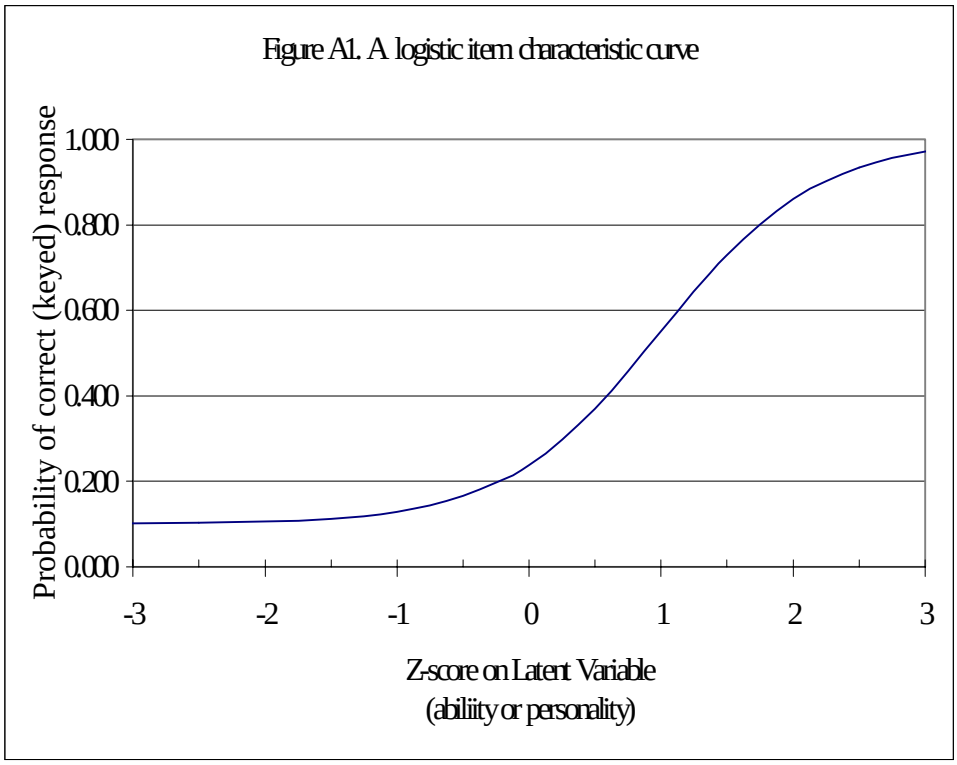
The following terms will help you understand DIF. In the following table, I adapt terminology discussed in Zumbo, Pope, Watson, & Hubley (1997):

Table A1.
Interpretation of ICC Properties for Cognitive and Personality Measures

ICC Property	**Cognitive, Aptitude, Achievement, or Knowledge Test**	**Personality, Social, or Attitude Measures**
Position along the X-axis (commonly called the b-parameter in IRT)	Item difficulty Amount of a latent variable needed to get an item right	Threshold Amount of a latent variable needed to endorse the item
Slope (commonly called the a-parameter in IRT)	Item discrimination. A flat ICC does not differentiate among test-takers	Item discrimination. A flat ICC does not differentiate among test-takers
Y-Intercept (commonly called the c-parameter in IRT)	Guessing	The likelihood of indiscriminate responding or social desirable responses

All of these parts of an ICC provide detailed information on various aspects of the item. Figure A1 gives an example of a parametric ICC. The horizontal axis is the continuum of variation for the latent variable. The scale of the latent variable is in z-scores. The item depicted in Figure A1 has a discrimination (i.e., slope) of 2.0, difficulty (i.e., threshold) of 0.60, and guessing parameter of 0.10.

Note that the item discrimination parameter determines how rapidly the curve rises from its lowest value of *c*, in this case 0.10, to 1.0. Note that if the curve is relatively flat then the item does not discriminate among individuals with high, moderate, or low total scores on the measure. Item discrimination values of 1.0 or greater are considered very good. Finally, note that the threshold parameter is the latent variable value on the continuum of variation at which the curve is midway between the lowest value, *c*, and 1.0, and therefore for achievement or aptitude measures, is a marker of the item difficulty. Items with difficulty values less than -1.0 indicate a fairly easy item whereas items with difficulty greater than 1.0 indicate rather difficult items.

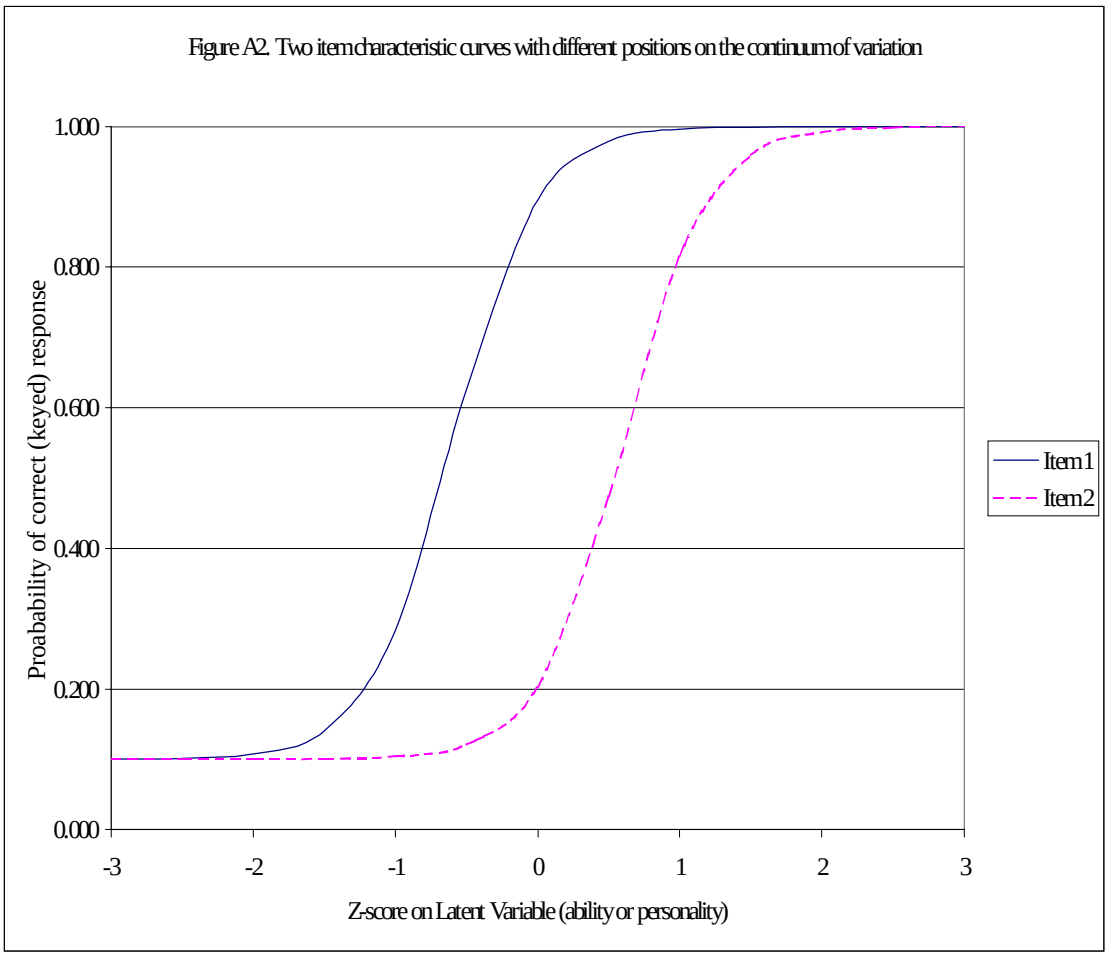


The ICCs shown in Figure A2 portray two items with equal slope (i.e., equal discrimination among respondents) but different placements on the continuum of variation. One would need to have more of the latent variable to endorse the item depicted by the dashed line than by the solid line. The dashed line is further to the right on the continuum. The dashed line, item 2, is thus considered more difficult.

In summary then, the ICC depicts the non-linear regression of the probability of the correct (keyed) response onto the continuum of variation conceptualized as the latent variable of interest. If we take an aptitude test as an example,

- the y-intercept is the likelihood of someone of very low aptitude getting the item correct,
- the slope of the curve indicates how well the item discriminates among levels of ability (a flat curve, for example, would mean that irrespective of the level of ability, individuals have the same likelihood of getting an item correct),

- the threshold indicates that value on the continuum of variation (the X-axis) at which the probability of getting the item correct is midway between the lowest value, *c*, and 1.0. For items for which the guessing parameter is zero, this midway value of course represents the value of the latent variable at which the likelihood of a correct response is greater than 0.50.



The continuum of variation has revolutionized psychometrics because traditional classical test theory methods are summary omnibus statistics that are, in essence, averages across the continuum of variation. For example, item total correlations (a traditional discrimination statistic) or reliability coefficients are one number irrespective of the level of individual variation. That is, the coefficient alpha is thought of as the same number irrespective of whether the respondent scored 3 standard deviations below, at, or 3 standard deviations above, the mean.

Therefore, these summary measures describe the sample as whole, ignoring how the psychometric properties of the scale may vary as a function of variation within the sample. Modern test theory builds on classical test methods and takes into account this sample variation and continuum of variation.

One natural psychometric technique that has arisen in this tide of modern psychometrics is that of DIF. Keeping in mind the continuum of variation, conceptually, DIF is assessed by

comparing the ICCs of different groups on an item. You can imagine that the same item is plotted separately for each group that the researcher wishes to evaluate (e.g., gender).

If the ICCs are identical for each group, or very close to identical, it can be said that the item does not display DIF. If, however, the ICCs are significantly different from one another across groups, then the item is said to show DIF. In most contexts, DIF is conceived of as a difference in placement (i.e., difficulty or threshold) of the two ICCS but as you will see in a few moments this does not necessarily have to be the case. Some examples of ICCs that do demonstrate DIF and some examples of items that do not demonstrate DIF will now be presented. Figure A3 is an example of an item that does not display DIF. As you can see, the area between the curves is very small and the parameters for each curve would be nearly equivalent.

Figure A3. An example of an item that does not display DIF.

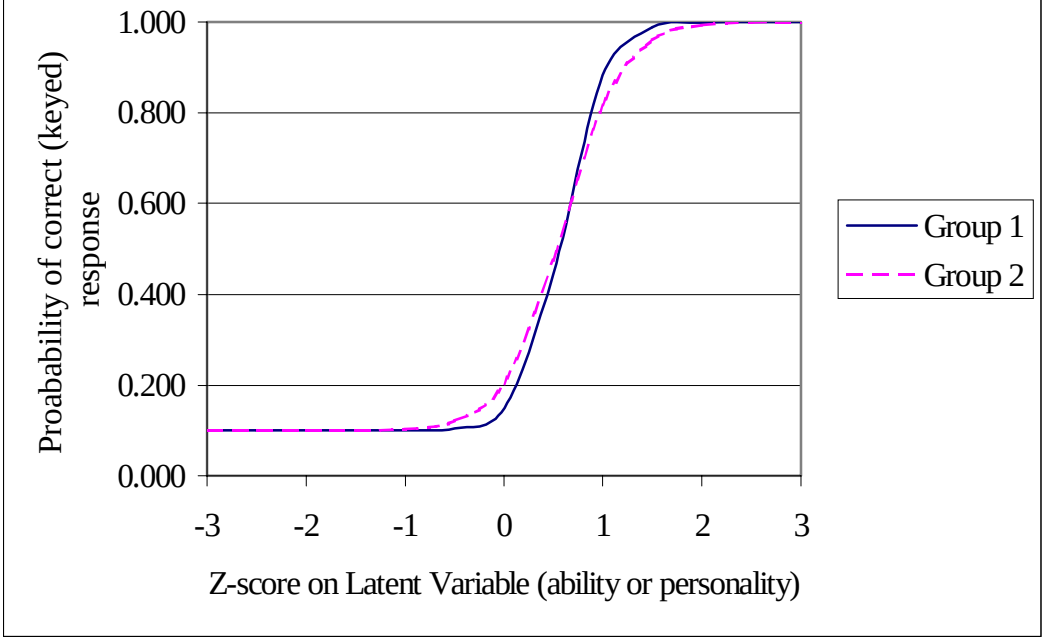


Figure A4, on the other hand, gives an example of an item that displays substantial DIF with a very large area between the two ICCs. This type of DIF is known as uniform DIF because the ICCs do not cross. An item such as the one shown in Figure A4 may not be an equivalent measure of the same latent variable for both groups.

Figure A4. An example of an item that displays substantial uniform DIF

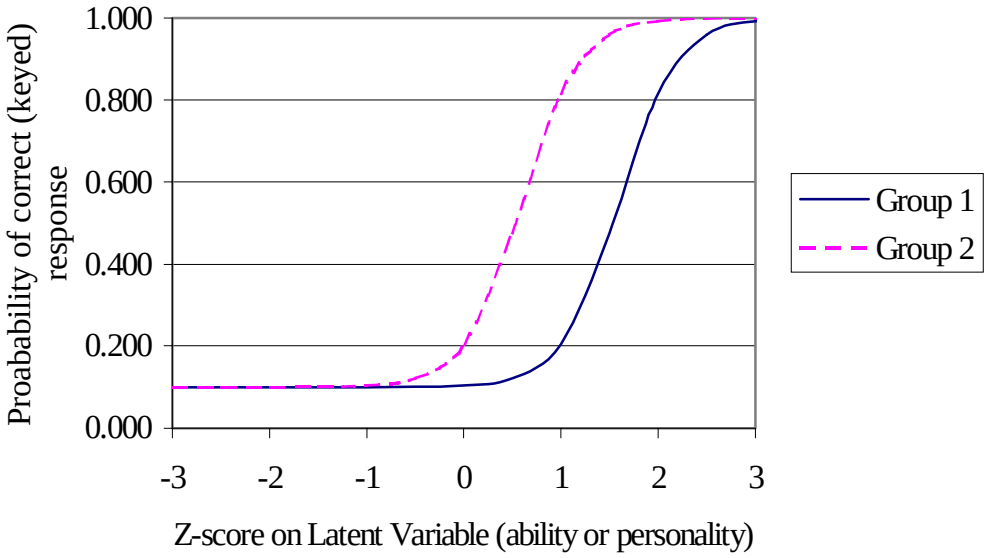
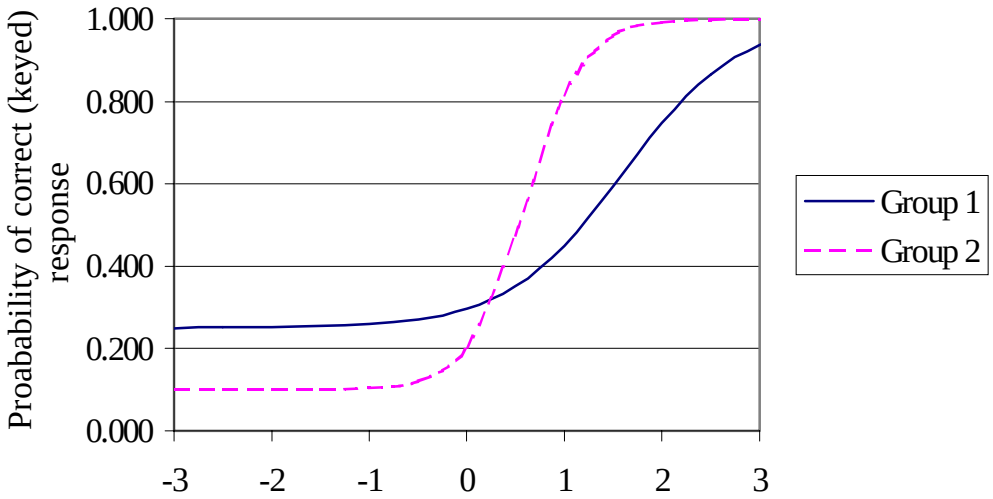


Figure A5 is an example of an item that displays substantial nonuniform DIF (i.e., the ICCs cross over one another). It depicts non-uniform DIF because for those individuals who score at or below the mean (i.e., $z \leq 0$), Group 1 is favored whereas for those scoring above the mean (i.e., $z>0$) Group 2 is favored. It would be an understatement to say that the example in Figure A5 is a rather complex (and exaggerated) form of DIF; however, DIF which depends on where you score on the continuum (not to the degree depicted in Figure A5) does periodically arise in practice.

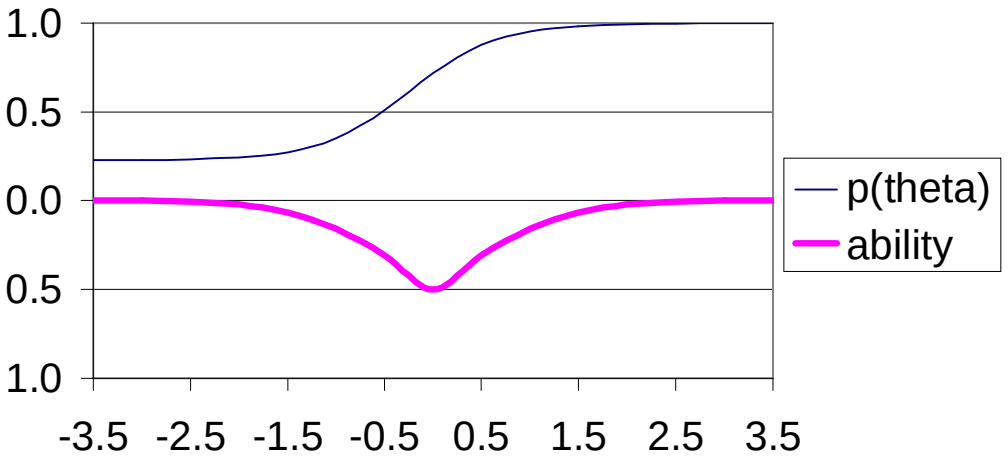
Figure A5. An example of an item that displays substantial nonuniform DIF



The Latent Distribution

To this point, we have not depicted what the latent distribution looks like in the IRT framework. Figure A6 an item characteristic together with the distribution of ability (or personality, or proficiency) for one group of respondents. Note that the latent ability distribution is Normal. Figure A7 depicts an item characteristic curve together with the distributions of ability for two groups of respondents. In this case, the ability distributions have equal variances by differ in the location of their means. If the IRT curves in Figure A7 differed then this may be seen as depicting impact.

Figure A6. An ICC with the latent distribuiton



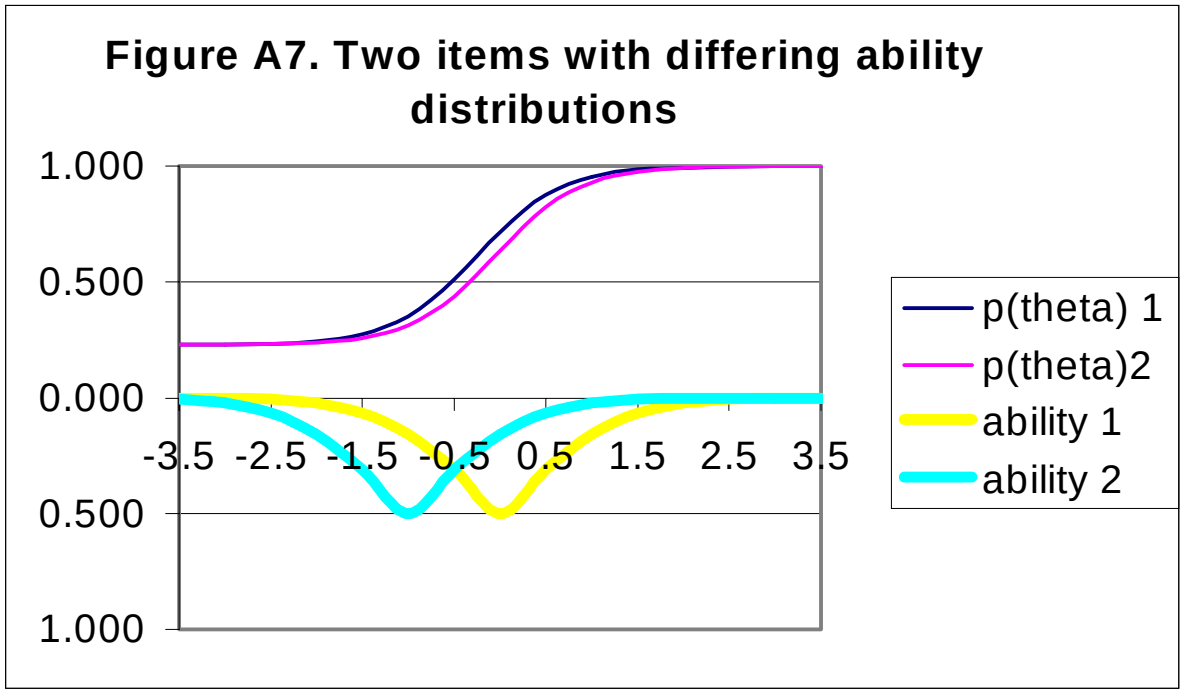


Figure A7. Two items with differing ability distributions
1.000
0.500
0.000
0.500
1.000
-3.5 -2.5 -1.5 -0.5 0.5 1.5 2.5 3.5
p(theta) 1
p(theta)2
ability 1
ability 2

APPENDIX B

Item parameters used in the simulation (38 items)

Item	a-parameter	b-parameter	c-parameter
1	.68060	-2.49389	.13492
2	1.38169	-1.08244	.42248
3	.53538	-2.16955	.18192
4	.86760	-.94422	.13269
5	1.06729	-.80361	.30321
6	1.10583	-.59550	.24694
7	1.53581	-.24833	.36850
8	1.46810	-.03225	.44504
9	1.04135	-.32195	.27445
10	1.29856	-.25170	.22733
11	.63728	-.40013	.27402
12	.72065	.20983	.18622
13	1.09920	.92697	.28737
14	.93156	1.53742	.21683
15	1.04698	-1.69322	.02896
16	.94893	-.80396	.22376
17	.69902	-.72047	.17965
18	1.11744	.26024	.32419
19	.82899	.11840	.35003
20	.68961	-.52557	.15307
21	1.04399	-.69906	.11490
22	1.26267	-.49348	.11900
23	.74013	-.02108	.18677
24	1.04001	.18751	.15927
25	.70804	-.01063	.12911
26	1.51116	.42848	.39223
27	.68581	-.30454	.10480
28	1.18653	.73290	.44767
29	.56387	.01254	.12300
30	.90872	.58845	.36072
31	1.12617	.08606	.21672
32	.86403	.37671	.24050
33	1.09621	.74282	.33839
34	.80390	.46551	.21272
35	.86075	.42113	.13494
36	.59933	.06103	.15025
37	.89227	1.46750	.26262
38	1.88982	.94262	.11904

Test-retest reliability assesses the degree to which test scores are consistent from one test administration to the next. Measurements are gathered from a single rater who uses the same methods or instruments and the same testing conditions.[4] This includes intra-rater reliability. Inter-method reliability assesses the degree to which test scores are consistent when there is a variation in the methods or instruments used.Â In practice, testing measures are never perfectly consistent. Theories of test reliability have been developed to estimate the effects of inconsistency on the accuracy of measurement. The basic starting point for almost all theories of test reliability is the idea that test scores reflect the influence of two sorts of factors:[7]. Assessing shy symptoms via computerized adaptive testing (CAT) provides greater measurement precision coupled with a lower test burden compared to conventional tests. The computerized adaptive test for shyness (CAT-Shyness) was developed based on a large sample of 1400 participants from China.Â CAT simulations based on the real data were carried out to investigate the reliability, validity, and predicted utility (sensitivity and specificity) of the CAT-Shyness. The CAT-Shyness item bank was successfully built and proved to have excellent psychometric properties: high content validity, unidimensionality, local independence, and no DIF. Testing and Assessment - Understanding Test Quality-Concepts of Reliability and Validity.Â The larger the reliability coefficient, the more repeatable or reliable the test scores. Table 1 serves as a general guideline for interpreting test reliability. However, do not select or reject a test solely based on the size of its reliability coefficient. To evaluate a test's reliability, you should consider the type of test, the type of reliability estimate reported, and the context in which the test will be used. Table 1. General Guidelines for. Reliability coefficient value. Another issue concerning reliability is that throughout developing tests, care should be exerted to make sure that the tests assess the actual skill of the test takers, systematic errors, and the unsystematic errors have no effect on test performances (Alderson et.al., 2005), to reduce or eliminate measurement error.Â Cover areas such as the testing situation, the test rubric, input the expected response and at the end, the relationship between input and response are covered by test method facets. Personal attributes encompass age, gender, cognitive style and background; later on, he lists a variety of random factors like tiredness, emotional condition and even some random differences in the testing context. Classical probability is the statistical concept that measures the likelihood of something happening. The chances of happening are equal.Â These examples include flipping coins, drawing cards from a deck, guessing on a multiple choice test, selecting jellybeans from a bag, and choosing people for a committee, etc. Classical Probability cannot be used: Dividing the number of events by the number of possible events is very simplistic, and it isnâ€™t suited to finding probabilities for a lot of situations. For example, natural events like weights, heights, and test scores need normal distribution probability charts to calculate probabilities. In fact, most â€œreal lifeâ€ things arenâ€™t simple events like coins, cards, or dice. Youâ€™ll need