Running head: A PSYCHOMETRIC ANALYSIS OF THE HOOKED ON NICOTINE CHECKLIST

A PSYCHOMETRIC ANALYSIS OF THE HOOKED ON NICOTINE CHECKLIST (HONC) WITH AN EYE TOWARDS GENDER DIFFERENTIAL ITEM FUNCTIONING: A CASE STUDY IN MISSING DATA AND DIF

by

CORNELIA ZEISSER

B.A. (Hons.), Simon Fraser University, 2002

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF EDUCATIONAL AND COUNSELING PSYCHOLOGY, AND SPECIAL EDUCATION

With Specialization in

MEASUREMENT, EVALUATION AND RESEARCH METHODOLOGY

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August 2004

Library Authorization

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

CORNELIA ZEISSER
Name of Author *(please print)*

03/09/2004
Date (dd/mm/yyyy)

Title of Thesis: A psychometric analysis of the Hooked on Nicotine Checklist (HONC) with an Eye towards gender differential item functioning: A case study in missing data and Dif

Degree: MA MERM Year: 2004

Department of ECPS
The University of British Columbia
Vancouver, BC Canada

Abstract

This study investigated gender Differential Item Functioning (DIF) in a measure of nicotine dependence (ND) in adolescents, the Hooked on Nicotine Checklist (HONC). First, a statistical modeling technique based on binary logistic regression was used to determine the presence and extent of DIF for each HONC item. Second, graphical DIF analyses were performed using nonparametric item response theory (NIRT). To investigate the impact of missing data on findings of DIF, all DIF analyses were performed on four different versions of the data: a) listwise deletion of missing cases (no imputation), b) imputation of missing values by row mode, c) imputation of missing values by column mode and d) imputation of missing values using the expectation maximization (EM) algorithm based on maximum likelihood estimation. Using binary logistic regression analysis, none of the ten HONC items were flagged as displaying DIF, with identical results across all four versions of the data. Using NIRT for graphical displays of DIF, seven out of ten HONC items showed DIF under column-wise and EM imputation of missing values, while eight out of ten HONC items were flagged as DIF under no imputation and row-wise imputation of missing values. The study concluded that missing data techniques did not have a strong influence on finding DIF. However, the importance of conducting DIF analyses with various DIF methods, including graphical displays, is emphasized, as the items displayed no DIF under logistic regression analyses, but marginal to substantial DIF using the NIRT. To potentially improve the HONC, it appears worth considering other possible dimensions to be included in the HONC as a measure of adolescent tobacco dependence pertaining to the psychological-social aspects of tobacco dependence in this particular population.

Table of Contents

Abstract ii
List of Tables v
List of Figures vii
Acknowledgements viii
Thesis Format 1

Chapter I: Introduction
1.1. Introduction 2
1.2. Study Purpose 2

Chapter II: Literature Review
2.1. Definition of Nicotine Dependence 4
2.2. Adolescent Smoking and Nicotine Dependence: Gender Differences Reported in the Literature 4
2.3. Focus on the Hooked On Nicotine Checklist (HONC): Development and Scoring 6
2.4. Reliability Studies on the HONC 8
2.5. Evidence Supporting the Validity of inferences made from the HONC 8
2.6. Factor Analysis of the HONC 9
2.7. Psychometric Performance Issues of the HONC and Study Rationale 9
2.7.1. Item Level Analyses of the HONC 13
2.7.1.1. Differential Item Functioning 13
2.7.1.2. Nonparametric Item Response Theory (NIRT) 14
2.8. Research Questions 14

Chapter III: Methodology
3.1. Respondents 15
3.2. Instruments 15
3.3. Analysis 15
3.3.1. Preparing Data Sets for Analysis Using Various Missing Data Techniques 15
3.3.2. Determining the Extent of Missing Data 18
3.3.3. Maximum Likelihood Estimation of Missing Data Using the EM Algorithm 18
3.3.4. Data Analysis 19
3.3.4.1. Scale Level Analyses: Factor Structure of the HONC 19
3.3.4.2. Item Level Analysis A: Differential Item Functioning (DIF) Methods Based on Binary Logistic Regression Analyses After Matching Respondents on the HONC Total Score 20
3.3.4.3. Item Level Analysis B: Graphical Representation of DIF Using Nonparametric Item Response Theory (NIRT) 23

Chapter IV: Results
4.1. Extent of missing data in the sample of adolescents who responded to the HONC within the British Columbia Survey of Smoking and Health.....26
4.2. Scale level analysis: Dimensionality of the HONC as determined by factor analysis for binary scored items.....30
4.2.1. Data set without imputation of missing values.....30
4.2.2. Data set with imputation of missing values by column mode value.....30
4.2.3. Data set with imputation of missing values by row mode value.....31
4.2.4. Data set with imputation of missing values using the EM algorithm.....31
4.3. Differential Item Functioning (DIF) based on binary logistic regression analysis after matching on the HONC total score.....33
4.3.1. Data set without imputation of missing values.....33
4.3.2. Data set with imputation of missing values by column mode value.....35
4.3.3. Data set with imputation of missing values by row mode value.....37
4.3.4. Data set with imputation of missing values using the EM algorithm.....39
4.4. Graphical Representation of DIF Using Nonparametric Item response Theory (NIRT): DIF assessed with TESTGRAF.....46
4.4.1. Data set without imputation of missing values.....46
4.4.2. Data set with imputation of missing values by column mode value.....49
4.4.3. Data set with imputation of missing values by row mode value.....49
4.4.4. Data set with imputation of missing values using the EM algorithm.....49

Chapter V: Discussion
5.1. Differential item functioning as assessed by statistical modeling using logistic regression.....65
5.2. Differential item functioning as determined by TestGraf.....68
5.3. Impact of missing data on findings of DIF.....69
5.4. Implications: What this study adds.....71
Conclusions.....72

References.....74

Appendices
Appendix A: Missing Data.....78
Appendix B: Factor Analysis of Binary Scored Items Based on the Tetrachoric Correlation Matrix.....86
Appendix C: Differential Item Functioning (DIF).....93
Appendix D: Nonparametric Item Response Theory (NIRT).....97

List of Tables

Table 1: Symptoms of Nicotine Dependence of 513 adolescents.....27

Table 2: Proportion of missing data for each HONC item in total, by gender and χ^2.....28

Table 3: Frequencies of missed item responses for males, females and in total.....29

Table 4: Factor loadings of the ten HONC items for the four versions of the data.....32

Table 5. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all HONC items without imputation of missing values.....34

Table 6. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for column mode-imputation of missing values.....36

Table 7. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for row mode-imputation of missing values.....38

Table 8. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for EM imputation of missing values.....40

Table 9: DIF results of logistic regression analysis without imputation of missing values: Model fitting Chi-square values, df and Nagelkerke R^2 values for items 4, 9 and 10.....42

Table 10: DIF results of logistic regression analysis with column-wise imputation of missing values: Model fitting Chi-square values and Nagelkerke R^2 values for items 4, 9 and 10.....43

Table 11: DIF results of logistic regression analysis with row-wise imputation of missing values: Model fitting Chi-square values and Nagelkerke R^2 values for items 4, 9 and 10.....44

Table 12: DIF results of logistic regression analysis with imputation of missing values using the EM-algorithm: Model fitting Chi-square values and Nagelkerke R^2 values for items 4, 9 and 10.....45

Table 13: TestGraf Composite DIF values for the HONC items in all versions of the Data.....47

Table 14: Comparison of logistic regression (logR) and TestGraf results (β) across different methods of missing value imputation.....48

List of Figures

Figure 1: The Hooked On Nicotine Checklist (HONC).....7

Figure 2: Analysis plan.....17

Figure 3: Plots of the TestGraf ICCs of all ten HONC items for the EM imputed data version.....51

Figure 4: Expected score plot for the EM imputed version of the data.....56

Figure 5: Plots of the TestGraf ICCs of all ten HONC items for the data version with column-mode imputation.....58

Figure 6: Expected score plot for the column-mode imputed data set.....63

Acknowledgments

I express my sincere thanks to my advisor, Professor Bruno D. Zumbo, for his boundless support and generous advice. I further thank Professor Anita Hubley and Dr. Susan James for their advice. I thank Professors Pamela Ratner, Joy Johnson and Joan Bottorff from the School of Nursing at the University of British Columbia for providing the data from the British Columbia Youth Survey on Smoking and Health. This study was supported by funding from the Social Science and Humanities Research Council (SSHRC).

Thesis format

This thesis is written in a format similar to an article reporting empirical findings in the measurement literature. Appendices are provided to give details that would not typically appear in a report on empirical findings. This format conforms to the University of British Columbia, Faculty of Graduate Studies, policy that states that theses may be written in article format.

Chapter I

1.1. Introduction

Many adolescents become dependent on nicotine but attempts at smoking cessation often fail. The Hooked On Nicotine Checklist (HONC) is widely used to measure nicotine dependence (ND) in adolescents (DiFranza et al., 2002a). The measure assesses typical smoking symptoms indicative of ND.

Although the psychometric properties of the HONC have been investigated (O'Loughlin et al., 2002a), a major issue in its application in research and practice is whether the items on the HONC are functioning the same for male and female adolescent respondents. Studies have consistently reported gender differences in self-reports of ND in adolescents (DiFranza et al, 2002a, DiFranza et al., 2002b, Psujek, Martz, Curtin, Michael & Aeschleman, 2004). However, there appears to be a lack of literature inquiring about the nature of these gender differences. As the HONC is widely used to assess ND in adolescents and to make decisions about intervention and smoking cessation strategies for this population, it is necessary for researchers, practitioners and policy makers to rule out test bias when using results based on the HONC. The next section will describe the study purpose and set the stage for this thesis.

1.2. Study Purpose

With an eye toward ruling out gender item bias, this thesis will begin with an overview of the literature on smoking and ND in adolescents. The main purpose of the present study is to investigate whether gender differences in self-reports of ND as assessed by the HONC are due to 'true' differences in the construct of ND between female and male adolescents, as opposed to a result of measurement artifact due to

differential item functioning (DIF) (Zumbo, 1999). An argument will be formulated that it is necessary to conduct item level analyses to determine the nature of the gender differences in adolescent ND reported in the smoking literature. This argument stresses that DIF studies for the HONC items are necessary and suitable to add more detailed psychometric knowledge about this instrument. New insights as to HONC items are functioning for males and females will aid researchers and practitioners alike in making informed and valid decisions about smoking intervention and cessation for adolescents based on the this measure of ND. Along the way, the commonly encountered problem of missing data will be addressed using various methodologies.

Chapter II

Literature Review

2.1. Definition of Nicotine Dependence: DSM-IV criteria and DiFranza's broader concept of ND based on addiction theory.

The construct of ND is generally defined as the compulsive use of cigarettes to achieve pleasurable effects and to avoid withdrawal symptoms (Fagerstrom & Schneider, 1989). According to the DSM-IV (APA, 1994), ND symptoms include a persistent desire to smoke and unsuccessful attempts to cut down or control usage of nicotine. Further symptoms of ND as described by the DSM-IV are social disruption caused by use and continued use despite physical or psychological symptoms caused by use (APA, 1994).

DiFranza et al. (2002a), however, note that these concepts of ND are too narrow for the context of adolescent smoking and suggested a broader conceptualization of ND in regards to adolescents. Thus, DiFranza et al. formulated the loss of autonomy theory by postulating that "a person is hooked when he/she has experienced a loss of autonomy over their use of nicotine" (DiFranza et al, 2002a). Based on the loss of autonomy theory, DiFranza et al. (2002a) suggest that ND occurs when autonomy over tobacco use is lost due to pharmacological, behavioral and psychological processes.

2.2. Adolescent smoking and nicotine dependence: gender differences reported in the literature.

The literature on self-reported ND in adolescents consistently reports gender differences. For example, DiFranza et al. (2002a) examined reports of the cumulative symptoms of ND in adolescents who had tried tobacco. These reports were based on the ten HONC items representing the symptoms of ND. The percentage of girls reporting

symptoms of ND was significantly higher than that of boys for seven of the ten HONC items. The study concluded that girls tend to develop symptoms of ND more rapidly than boys (DiFranza et al, 2002a).

In a similar vein, the DANDY (Development and Assessment of Nicotine Dependence in Youths) study (DiFranza et al., 2002b) reported gender differences in how adolescents reported symptom onset of ND. HONC mean scale scores were compared between two groups: adolescents who had tried tobacco and adolescents who were monthly smokers. In both groups, HONC mean scale scores were higher for females than for males (DiFranza et al., 2002b). A limitation, in terms of generalizability, of these two studies was the narrow age range of subjects, as they were all seventh grade students age 12-13 years. The HONC, as a diagnostic measure of ND, can be expected to perform differently depending on the characteristics of the population studied. That is, different results might be obtained in adolescents older or younger than those in these particular studies.

Psujek et al. (2004) raised the issue that gender differences in self-reported ND may be apparent due to how ND has been operationalized. This point is of importance for the psychometric analysis of the HONC proposed here. Even though Psujek et al.'s study used measures other than the HONC, this point helps make the case for a thorough investigation of item bias and measurement artifact in the HONC by conducting DIF analyses. That is, DIF analyses of measures of ND in adolescents help clarify the definition and operationalization of ND by systematically ruling out item bias with respect to gender in the measures used.

2.3. Focus on the Hooked on Nicotine Checklist (HONC): Development and scoring.

The HONC was developed using adolescent subjects. The measure was derived from the loss of autonomy theory discussed in the previous section of the literature review (DiFranza et al., 2002a). Based on the notions of this theory, the endorsement of a single HONC item indicates loss of autonomy over nicotine use (DiFranza et al., 2002a) and is associated with a failed attempt at cessation (O'Loughlin et al., 2002a). An example item is "Do you smoke because it is really hard to quit". Item responses are in binary format (Yes=1/No=0) and are summed for a cumulative HONC scale score presumed to reflect loss of autonomy over tobacco use (DiFranza et al., 2002a). Figure 1 shows the ten items of the HONC.

Figure 1: The Hooked on Nicotine Checklist (HONC).

1. Do you smoke because it is really hard to quit? (Y/N)
2. Have you ever felt like you were addicted to tobacco? (Y/N)
3. Is it hard to keep from smoking in places where you are not supposed to (i.e., at school)? (Y/N)

When you tried to quit smoking or when you haven't use tobacco in a while, did you or do you :

4. Find it hard to concentrate? (Y/N)
5. Feel more irritable? (Y/N)
6. Feel a strong urge to smoke? (Y/N)
7. Feel restless? (Y/N)
8. Feel sad, blue or depressed? (Y/N)
9. Feel stressed? (Y/N)
10. Feel light-headed? (Y/N)

2.4. Reliability studies on the HONC

O'Loughlin, Tarasuk, DiFranza & Paradis (2002c) produced test-retest reliabilities ranging from $\kappa = 0.61 - 0.93$ for nine of the ten dichotomously scored HONC items. Only one item tapping into depression (HONC item 8 - feel depressed when tried to quit) showed poor test-retest agreement with a Kappa coefficient of 0.34 (O'Loughlin et al., 2002c). Further, a limitation of the HONC with respect to its internal consistency is that HONC item 8 had a low item-total correlation of 0.19, indicating that this HONC item is not consistent with the underlying construct presumably measured by the checklist. Despite this one item showing a low item-total correlation, high internal consistency was demonstrated for the HONC by O'Loughlin et al. (2002c) with Cronbach's $\alpha = 0.90$. However, a limitation of the latter study is that results were based on a convenience sample, which may not be representative of the larger population of adolescents. A different sample with less variability could also yield a lower coefficient alpha. Further, the above study could not assess criterion related validity of the HONC, as a "gold standard" for the validity of inferences made from the HONC is yet to be established.

2.5. Evidence supporting the validity of inferences made from the HONC

Evidence in support of content validity of inferences based on the HONC was provided in a qualitative study of adolescent smokers' experience of ND (O'Loughlin, Kishchuk, DiFranza, Tremblay, Paradis, 2002b). That is, in focus group discussions, adolescent smokers endorsed all of the ND symptoms on the HONC as relevant to their experience, again with the exception of HONC item 8 ('feel sad, blue or depressed when tried to quit'). This is the same item that had previously shown low test-retest agreement and a poor correlation with the HONC total score in O'Loughlin's (2002a) study.

However, the strength of O'Loughlin et al.'s (2002b) study is that it demonstrated in qualitative terms that adolescents are able to provide self-reports of symptoms of ND that are consistent with the theoretically driven conceptualization of ND that forms the basis for the HONC. This study also highlighted the possibility of improving the psychometric properties of the HONC by eliminating the item related to feeling depressed during nicotine withdrawal (item 8).

Additional evidence for construct validity of the HONC is based on the finding that the HONC is highly correlated with the amount smoked and failed attempts at cessation (DiFranza, 2002b). Further, O'Loughlin et al. (2002a) demonstrated that the HONC showed high agreement with another indicator of ND, the ICD-10 (International Classification of Diseases, 10^{th} revision) criteria for tobacco dependence (κ=.74).

2.6. Factor analysis of the HONC.

The HONC putatively taps into three dimensions: self-medication, negative reinforcement and incentive sensitization (O'Loughlin et al., 2002b). However, DiFranza (2002b) reported that a one-factor solution fits the data well. This is an important psychometric property, as it lends justification to using a HONC total score in subsequent scale score interpretation and also in DIF analyses, because it suggests that the HONC items tap into one dimension – ND.

2.7. Psychometric performance issues of the HONC and study rationale

As indicated, the literature on adolescent smoking and ND points toward females reporting more symptoms of ND on the HONC than males. This leads to the question of whether females endorse items more frequently than males due to true differences in the construct measured (ND) as opposed to *due to HONC item characteristics* irrelevant to

the measurement purpose. Are HONC items functioning the same way for females as for males? Do gender differences on the HONC indeed reflect true differences in ND between males and females? There are many plausible explanations for why females endorse HONC items differently than males. For example, are items possibly measuring other dimensions besides the construct of interest (ND), such as social pressure or adolescent's desire to please/to impress others? Such dimensions could impact how female adolescents respond to the HONC items, as opposed to males. That is, are certain HONC items possibly tapping a dimension related to how females read certain statements about tobacco use that does not apply for males? It is generally known that smoking in adolescents is related to important developmental processes, such as constructing a social identity (Schillington & Clapp, 2000). Thus, as Johnson and colleagues (in press) state, adolescent tobacco dependence may be conceptualized as being multidimensional, including aspects such as social, pleasurable, empowering and emotional (Johnson et al., in press). These non-chemical aspects of smoking may possibly influence males' and females' responses to certain HONC items in different ways.

Another example pointing out an alternative explanation for why females score higher on self-reported ND is related to self-report stability with respect to substance use in general. For instance, perhaps gender differences in item responses are due to the propensity towards (dis)honsesty in self-reporting, particularly by males? A study on self-report stability by gender found that discrepant reports were highest among cigarette users (before alcohol and marijuana) with 18.5% reporting tobacco use at baseline but not at follow-up (Shillington, Cotler, Mager & Compton, 1995). Males were significantly more likely to be discrepant in their reports of cigarette use than females. These findings

further highlight the possibility that stability of adolescents' ND self-reports on the HONC may be similarly compromised, possibly due to item characteristics affecting females differently than males. One social aspect potentially playing a role in adolescents' differing endorsement of HONC items is that females could be more willing to report certain symptoms of ND as addressed by the HONC, while males could be more reluctant to report such symptoms. To summarize, the wide variety of social processes playing into tobacco use in adolescents makes it difficult to interpret current findings of gender differences in their self-reports of ND. Specifically, discrepancies of this kind may be DIF related. It is important to clarify the nature of such gender discrepancies in reporting because they pose a potential threat to the validity of inferences drawn from self-report data on ND in adolescents.

In order to tackle the difficulties in interpreting findings of gender differences, it is indicated to rule out measurement artifact due to item bias. For this purpose, one needs to begin by clarifying whether gender differences in responding to HONC items are a result of DIF and to what extent. Which items function differentially for females and males? To what extent are differences in item responses in self-reported ND for males and females related to properties of the item, rather than to true differences in the construct (ND)? Indeed, gender differences found on the HONC could be interpreted as 'true differences' in ND between males and females: different groups of respondents have differing probabilities of endorsing items due to true differences between females and males in the underlying construct being measured by the items. In this case, we would have item impact (Zumbo, 1999). However, without DIF analyses, researchers cannot be sure about this interpretation. Therefore, the question of interest to researchers must be to what

extent the HONC items display DIF. Although a substantial body of literature reports gender differences in adolescents' ND, (e.g., DiFranza et al, 2002a, DiFranza et al., 2002b, Psujek et al., 2004), no reference is available to researchers that specifically investigates the nature of such gender differences. In particular, thorough psychometric analyses of the gender differences on the HONC are absent from the literature. However, it is necessary to clarify psychometric performance issues of the HONC with respect to the gender differences reported in the literature. This question is best pursued at the item-by-item level, rather than at the scale level. That is, a mere examination of the full checklist, summing responses to a HONC total score, would not inform the researcher about gender DIF or item bias. The researcher would not be able to ascertain whether the results of gender differences are due to true differences in ND, as opposed to measurement issues above and beyond the construct of interest. Therefore, the present study aimed to produce new insights into the nature of the gender differences in responses to the HONC by applying two classes of item level analyses – Differential Item Functioning (DIF) and nonparametric item response theory (NIRT). These methods were used to examine to what extent the gender differences found using the HONC are due to measurement artifact. What follows is a brief description of how DIF analyses can provide important new insights into the psychometric performance of the HONC. The instrument should not measure other differences between male and female adolescents, apart from true differences in ND. The section concludes with a brief description of NIRT and how this technique is useful in determining on an item-by-item level if there is measurement artifact with respect to gender and ND.

2.7.1. Item Level Analyses of the HONC

2.7.1.1. Differential Item Functioning (DIF).

There are various reasons why DIF studies on the HONC are important. If there is DIF, then gender differences on the HONC may reflect item bias, rather than true differences in ND. Hence, it is necessary to identify and eliminate these measurement artifacts so that inferences made from HONC test scores are not compromised. The validity of self-report measures of ND is of great importance to researchers investigating adolescent smoking patterns and ND. Self-report measures are the most common source of information concerning adolescents' use of tobacco. Moreover, such self-reports form the basis for intervention and smoking cessation strategies.

Above and beyond existing psychometric knowledge about the HONC as a full checklist, DIF analyses add important information for researchers and practitioners. That is, analyses at the item level of the HONC help unravel complex issues underlying differential performance on the instrument for different groups of respondents (e.g., males and females). Thus, by providing information about psychometric performance at the item level, DIF analyses yield insights into broader measurement and validity issues about the HONC that still need to be established. For example, item level analyses are desired to obtain information about how HONC items perform at different levels of the construct (ND). Further, if an item displays DIF, then that item may be measuring a secondary nuisance factor differently for males and for females. Thus, DIF investigations provide important information for researchers and practitioners using the HONC to make more valid assessments of adolescents' levels of ND. For a more detailed discussion of

the logistic regression-based DIF method (Zumbo, 1999) used in the present study, please refer to Appendix B, Differential Item Functioning.

2.7.1.2. *Nonparametric Item Response Theory (NIRT).*

To this point, no published literature appears to exist involving IRT/NIRT studies of the HONC. However, a NIRT approach to the HONC is needed to obtain information at the item level about how the instrument performs at different levels of construct (ND). Such information cannot be provided by classical test theory (CTT) methods, such as reliability theory. Further, in contrast to parametric item response theory (PIRT), which demands large item pools and sample sizes for accurate parameter estimation, the main advantage of NIRT is that it works well for small sets of items and small sample sizes, such as those encountered in the present study. A more detailed discussion of the advantages of using NIRT techniques for investigating psychometric performance issues of the HONC is provided in Appendix D, Nonparametric Item Response Theory (NIRT).

2.8. Research Questions

Following the above study rationale, the primary research question of the present study was whether items on the HONC display gender DIF effects. A secondary research question was whether a finding of DIF is impacted by the extent of missing data or by the missing data technique used. This research question aims to explore how various missing data techniques may yield different results in terms of DIF findings in the HONC. The secondary research question posed in this investigation thus aims to increase understanding of appropriately applying missing data techniques, given the study context and extent of missing values in the data set at hand.

Chapter III

Methods

3.1. Respondents

The data were obtained with permission from the Nursing and Health Behaviour Research Unit (NAHBR) at the University of British Columbia (Johnson, Ratner & Bottorff, 2003). Respondents were 513 adolescents (256 males, 257 females) who completed the HONC within the British Columbia Youth Survey on Smoking and Health (BCYSOSH; conducted in 2001/02) as a self-administered paper-and-pencil questionnaire during class time. The adolescents ranged in age from 14 to 20 years. Their mean age overall was 16.0 years, with a standard deviation of 1.06. The mean age for males was 16.2 years, with a SD of 1.06. For females, the mean age was 15.9 years, with a SD of 1.04.

3.2. Instruments

The HONC was described in detail earlier in the literature review section in Chapter I of this manuscript. See also Figure 1 for a complete display of all 10 HONC items.

3.3. Analysis

3.3.1. Preparing data sets for analysis using various missing data techniques

To investigate the matter of missing data, four different data sets were created for analysis. However, prior to creating the four versions of the data, a decision rule was established that cases with more than three missed item responses out of the total of ten HONC items be deleted and excluded from the analyses. As a result, 89 cases were dropped. One data set was created choosing the SPSS default listwise deletion (LD). This approach is also referred to as complete case analysis, as only cases with no missing

values on any of the items are used. Thus, on this original version of the data, there was no imputation of missing values. Next, a data set was created in which single values were imputed by rows, replacing the missing value by the row mode value, i.e., subject mode across each item. The third data set was created by single imputation of the column mode value, i.e., item mode across subjects. The mode was chosen for single value imputation because the HONC items are in binary response format (0 representing "No" and 1 representing "Yes"). Finally, a fourth version of the original data was created containing estimates of missing values based on maximum likelihood estimation using the EM algorithm in the computer program PRELIS. Figure 2 depicts the various versions of the data analyzed, i.e. applying various methods of dealing with missing data at each level of the analysis.

Figure 2. Analysis Plan.

	No Imputation of Missing Values (Complete Case Analysis)	**Imputation by Row Mode Value**	**Imputation by Column Mode Value**	**Imputation Using the EM Algorithm**
Determining the Dimensionality of the HONC with FA of the Tetrachoric Correlation Matrix	Listwise deletion; use all of the items with complete information only in assessing dimensionality of the HONC	Use a complete data set, acknowledging that variability may have been lost due to row mode value imputation	Use a complete data set, acknowledging that variability may have been lost due to column mode value imputation	Use a complete data set containing estimates of missing values based on joint distributions of fully observed variables
DIF Analysis using Binary Logistic Regression	Listwise deletion; use all of the items with complete information only	As above	As above	As above
DIF Analysis using Nonparametric Item Response Modeling in TestGraf	As above	As above	As above	As above

3.3.2. Determining the extent of missing data

First, the extent of missing values in the data set overall was determined. Next, it was determined whether the extent of missing data varied by gender, that is, whether males or females skipped more item responses. HONC item endorsement frequencies were computed for the sample overall and by gender.

Four different approaches to handling the missing data problem were utilized, as described in more detail in Appendix A, Missing Data. In general, ad hoc methods for handling missing data, such as listwise deletion, pairwise deletion and mean/mode imputation are more widely known and are thus not discussed in detail. Maximum likelihood methods, however, are newer methods for dealing with missing data. One of these methods, the EM algorithm was used in this study and is therefore described in more detail. The next section briefly introduces this maximum likelihood method for the purpose of setting the stage for this procedure, while details are provided in Appendix A, Missing Data.

3.3.3. Maximum likelihood estimation of missing values with focus on the EM algorithm

Missing values were estimated by applying maximum likelihood estimation of missing data using the EM (expectation-maximization) algorithm (Enders, 2001; Pigott, 2001). Maximum likelihood methods for missing data are based on distributional models for the data (Pigott, 2001). Therefore, the joint distribution of all variables, including outcomes and predictors is required to be multivariate normal. Further, maximum likelihood methods assume an ignorable response mechanism for the missing data (see Appendix A: Missing Data). When data are missing at random (MAR) or missing

completely at random (MCAR), the response mechanism is deemed ignorable (Pigott, 2001). MAR describes data that are missing for reasons related to completely observed variables in the data set. In order for missing values to be MCAR, however, the reasons for the missing values must be unrelated to other variables in the data set. Given that these assumptions hold, maximum likelihood methods have the advantage of being appropriate for a wider range of situations than the above-mentioned ad hoc methods, such as complete case analysis (Pigott, 2001). Complete case analysis refers to an analysis in which only cases with no missing values are used; cases that have missing data on any of the variables are deleted. For a more detailed discussion of applying maximum likelihood estimation of missing data using the EM algorithm, see Appendix A, Missing Data. The next section will describe the two levels of analysis performed on the four different versions of the HONC data.

3.3.4. Data analysis

3.3.4.1. Scale level analysis: factor structure of the HONC.

Overall, two types of analyses of the HONC were conducted: a scale level analysis and an item level analysis. The scale level analysis is necessary for determining the factor structure of the HONC for the particular sample at hand. That is, at least essential unidimensionality is necessary to justify matching on the HONC total score in the subsequent DIF analyses. This would be the case if one dominant factor were found to be present using the Eigenvalue greater than one-rule, or if the ratio of first to second Eigenvalues is large enough to conclude that the first factor accounts for a substantial proportion of variance in the data, compared to any subsequent factors. The dimensionality of the HONC was examined applying principal components analysis

(PCA) and factor analysis for binary items using the MINRES (minimal residuals) procedure in the computer program PRELIS. As the HONC items are binary scored (Y/N), it was necessary to factor analyze a matrix of tetrachoric inter-item correlations, rather than the usual matrix of Pearson correlations. For a more detailed discussion of the issues around factor analysis of binary scored data, please refer to Appendix B: Factor Analysis of Tetrachoric Correlation Matrices.

Once the dimensionality of the HONC in the present sample was determined and matching on the HONC total score could be justified based on unidimensionality, the first set of item level analyses was conducted using the logistic regression-based DIF method developed by Zumbo (1999), as described in the next section.

3.3.4.2. Item level analysis A: Differential Item Functioning (DIF) methods based on binary logistic regression analysis after matching on the HONC total score.

At the item level, DIF was first investigated using statistical modeling techniques. DIF analyses developed by Zumbo (1999) were conducted to assess whether the HONC items function the same way for female and male respondents. As the HONC items are binary scored, binary logistic regression analysis was the statistical model of choice to investigate DIF for each of the HONC items (Clauser & Mazor, 1998).

Prior to investigating whether there is a gender effect, the groups (i.e., males and females) were matched on the variable of interest, the total scale score for the HONC. That is, there was statistical conditioning on the underlying trait that the items are intented to measure before group differences were investigated to determine whether the test items are problematic for any particular group of interest (Zumbo, 1999).

As indicated, the DIF analysis using statistical modeling was performed to measure if there is a gender difference for each HONC item, after matching on the total score. If this is the case, then there is DIF. The dependent variables are the item responses. The independent variables are the grouping variable (GRP), the HONC total score for each respondent (TOTAL), and the group by total interaction (Zumbo, 1999). Main effects of gender differences were studied to assess uniform DIF. Non-uniform DIF was studied by examining the gender by total score interaction. Using these variables, a linear regression equation can be stated regressing the independent variables on a latent continuously distributed random variable, y* (Zumbo, 1999). Thus, the model is defined as follows:

$$(1) \qquad y^* = b_0+b_1TOTAL+B_2GRP+B_3TOTAL*GRP_i$$

However, y* is the natural log of the odds ratio:

$$(2) \qquad \ln [p_i/(1-p_i)] = b_0+b_1tot+b_2group+b_3(tot*group),$$

where p is the proportion of respondents who endorse the item in the direction of the latent trait (Zumbo, 1999). The 2-degree of freedom Chi-Square test can be used to test for both uniform and non-uniform DIF. This model provides a DIF analysis conditionally on the relationship between the item response and the total score; the effects of group (i.e. gender) are tested for uniform DIF and the interaction of group and total score to assess non-uniform DIF (Zumbo, 1999). The advantage of using this logistic regression method

over other DIF methods, such as the Mantel Haenszel method, is that one can model uniform and non-uniform DIF simultaneously or independently (Swaminathan, 1994).

The statistical DIF analysis was conducted in three steps, entering the variables into the model above in a hierarchy. First, only the total score was examined for DIF. In the second step, the grouping variable (gender) was investigated. Finally, there was an investigation of the total by gender interaction for DIF. This step examines whether the difference between the group responses on an item varies over the latent variable continuum. From each of the three steps, a Chi-squared statistic was obtained. It was used in the statistical test for DIF.

Further, an effect size estimator R^2 was computed for each step (Zumbo, 1999). The magnitude of DIF can be computed by substracting the R^2 value of the first step from that of the third step. To identify an item as displaying DIF, both the Chi-squared significance and the corresponding effect size measure must be considered. Specifically, the 2-df Chi-squared test for DIF, testing for gender and interaction effects, must have a p-value less than or equal to 0.01. Further, the corresponding effect size measure must have an R^2 value of at least 0.035, following the criteria provided by Jodoin and Gierl (2001). According to these criteria, the magnitude of DIF was quantified as follows: R^2 values below 0.035 for negligible DIF, between 0.035 and 0.070 for moderate DIF, and greater than 0.070 for large DIF.

If DIF is found for an item, the steps in the computation of DIF can be used to determine if the DIF is uniform or non-uniform (Zumbo, 1999). Uniform DIF is determined by a comparison of the R^2 values between the first and second steps to measure the unique variation explained by the gender differences over and above the

conditioning variable (HONC total score). Now, let us turn towards the second type of item level analysis to assess DIF performed on the four versions of the HONC data.

3.3.4.3. Item Level Analysis B: Graphical Representation of DIF Using Nonparametric Item response Theory (NIRT).

In addition, DIF was assessed for each item through graphical representation. The visualization method for DIF is a nonparametric regression. The graphical display of DIF is based on investigating the relationship between total score and item responses for each group separately but on the same graph. This way, group differences in responding to the item can be visualized. The lines displayed on the graphic represent item response functions (IRFs). Differences between the two IRFs represent the group differences (i.e., males/females) in item responding to each individual HONC item. However, in the graphical approach, the total score is not held constant, as in the statistical approach (Zumbo, 1999). Instead, the graphical approach using TestGraf measures and displays DIF as a designated area between the curves of the two groups to determine the presence and extent of DIF. The area between the two IRFs is denoted as beta (β) and represents the amount of DIF. That is, if there is no area between the two curves, then there is no DIF. Conversely, if the area between the two lines is large, this indicates substantial DIF for this item. The area β measures the weighted expected score discrepancy between the two groups of examinees with the same underlying ability on a particular item. TestGraf computes the DIF summary index of β as follows:

$$\beta_{img}(\theta) = \sum_{q=1}^{Q} p_{Fq}\left[P_{im}^{(F)}(\theta) - P_{im}^{(R)}(\theta)\right],$$

where $P_{im}^{(R)}(\theta)$ and $P_{im}^{(F)}(\theta)$ stand for the option characteristic curve values for the reference and focal groups, respectively. In the above equation, *i* denotes item, while *m* denotes item option, and g denotes group. For all TestGraf analyses, the smoothing parameter for obtaining the ICCs was set to 0.45.

Next, the graphical displays of TestGraf were examined for each item, in each version of the data to determine whether the HONC items are functioning appropriately for both males and females and the extent to which each of the HONC items discriminates among respondents. For example, for an item functioning appropriately, one would expect the slope to begin at the bottom left and to steadily increase toward the top right of the ICC plot (Ellis, 1989). In order to decide whether an item displays DIF, cut-off values were used based on the cut-off indices produced by Zumbo and Witarsa (2003) for β in identifying TestGraf DIF for the sample size combination N_1/N_2 = 200/100 and a level of significance (α) of.01. The above sample size combination was chosen to reflect a more conservative approach, as the next higher sample size combination provided was N_1/N_2 = 500/500; in the present analysis, the actual sample size combination was N_1/N_2 = 257/256. The cut-off value used in all TestGraf DIF analyses was .0415. That is, any item displaying a composite DIF index of .0415 or greater was identified as DIF.

The use of the graphical approach may be viewed as complementary to the statistical approach to determining DIF. Thus, the two methods together can provide complementary evidence for the presence of DIF (Slocum, Gelin & Zumbo, 2003, in press).

For a more detailed discussion of NIRT, please consult Appendix D: Nonparametric Item Response Theory (NIRT).

Chapter IV

Results

4.1. Extent of missing data in the sample of adolescents who responded to the HONC within the British Columbia Survey of Smoking and Health

For descriptive purposes and comparison, Table 1 shows the endorsement frequencies for the 10 HONC items (symptoms of ND) in percentages and counts (in brackets) for the sample of 513 adolescents overall as well as by gender. Item-total correlations are also provided for each HONC item in the last column of the table.

Of all HONC items, item 3 showed the greatest extent of missing data, with 12.3% (N=63) missed item responses. This large proportion of missing data for this particular item is possibly due to how the question is worded, i.e. featuring the negative "not" as part of the question. Thus, some adolescents may have skipped this question because they were not certain as to its exact meaning. Table 2 provides further details as to the proportion of missing data for each of the ten HONC items in regards to the total sample, as well as a breakdown of missed item responses for each HONC item by gender. The χ^2 significance test for a gender effect in skipping responses was non-significant for all ten HONC items. Table 3 provides the item-by item frequencies of missed item responses for males, females and in total.

In summary, the extent of missing data on this sample was consistent across the ten items of the HONC, with proportions of missed item responses mostly ranging from 1.6 to 2.7 % overall. As indicated, an exception was item 3 (Is it hard to keep from smoking in places where you are not supposed to (i.e. at school?). This item was missed approximately five times as often as any other HONC item.

Table 1. Symptoms of Nicotine Dependence as endorsed by 513 adolescents who responded to the HONC within the British Columbia Survey on Smoking and Health

HONC items/ Symptoms of ND	**Percentage of overall sample reporting symptom (n)**	**Percentage of 256 boys reporting symptom (n)**	**Percentage of 257 girls reporting symptom (n)**	**Item-Total correlation**
1. Do you smoke because it is really hard to quit?	36 (181)	34 (87)	37 (94)	.659
2. Have you ever felt like you were addicted to tobacco?	57 (287)	53 (130)	62 (157)	.733
3. Is it hard to keep from smoking in places were you are not supposed to (i.e. at school)?	48 (216)	45 (100)	51 (116)	.552
When you tried to quit smoking or when you haven't use tobacco in a while, did you or do you:				
4. Find it hard to concentrate?	43 (218)	43 (107)	44 (111)	.775
5. Feel more irritable?	52 (261)	44 (110)	61 (151)	.795
6. Feel a strong urge to smoke?	64 (319)	56 (139)	72 (180)	.773
7. Feel restless?	46 (230)	38 (96)	54 (134)	.750
8. Feel sad, blue or depressed?	31 (153)	24 (61)	37 (92)	.581
9. Feel stressed?	59 (292)	49 (123)	68 (169)	.730
10. Feel light-headed?	17 (84)	7 (18)	16 (40)	.386

Coefficient alpha: .87 (n = 424)
Standardized item alpha: .87

Table 2: Proportion of missing data per HONC item in total and by gender

HONC Item	**Missing in Total**		**Missing by Gender**			
	%	**(n)**	**Male**		**Female**	
			%	**(n)**	**%**	**(n)**
HONC item1	1.6	(8)	1.2	(3)	1.9	(5)
HONC item2	2.3	(12)	3.5	(9)	1.2	(3)
HONC item3	12.3	(63)	12.9	(33)	11.7	(30)
HONC item4	1.9	(10)	1.6	(4)	2.3	(6)
HONC item5	2.5	(13)	2.0	(5)	3.1	(8)
HONC item6	2.5	(13)	2.7	(7)	2.3	(6)
HONC item7	2.7	(14)	2.0	(5)	3.5	(9)
HONC item8	2.5	(13)	2.3	(6)	2.7	(7)
HONC item9	2.7	(14)	2.0	(5)	3.5	(9)
HONC item10	2.3	(12)	2.0	(5)	2.7	(7)

Table 3: Frequencies of missed item responses for males, females and in total

Number of Missed item Responses	**Gender**				**Total**	
	Male		**Female**			
	%	**(n)**	**%**	**(n)**	**%**	**(n)**
0	80.9	(207)	84.4	(217)	82.7	(424)
1	16.0	(41)	11.7	(30)	13.8	(71)
2	1.2	(3)	1.2	(3)	1.2	(6)
3	.4	(1)	0	(0)	.2	(1)
4	0	(0)	0	(0)	0	(0)
5	0	(0)	0	(0)	0	(0)
6	0	(0)	0	(0)	0	(0)
7	.4	(1)	1.6	(4)	1.0	(5)
8	.8	(2)	.4	(1)	.6	(3)
9	.4	(1)	.8	(2)	.6	(3)

4.2. Scale level analysis: Dimensionality of the HONC as determined by factor analysis for binary scored items

The following section provides the results of the PCA for all four versions of the data, based on the Eigenvalue greater than one-rule, as well as the results of the factor analysis using the minimal residual procedure (MINRES). The reader is reminded that, before the analysis proceeded, a rule was established that a participant may have missed no more than three out of the 10 HONC items; 11 cases (2.2 %) with more than 3 missed item responses were not used.

4.2.1. Data set without imputation of missing values.

The total effective sample size for the analysis of this first version of the data was N=424. A dominant first principal component was shown, with an Eigenvalue of 6.53, explaining 65.3 % of the total variance. Thus, it is justified to conclude that the HONC is strongly unidimensional for this version of the data. The factor loadings for the dominant first factor ranged between 0.468 and 0.924. For factor loadings of individual HONC items for all four versions of the data, please refer to Table 4.

4.2.2. Data set with imputation of missing values by column mode value.

For the other versions of the data, that is, applying various missing data techniques, the results were very similar. The total effective sample size for the version created by imputing the column mode value was N=502. The Eigenvalue of the first principal component was 6.38, explaining 63.8 % of the total variance. Therefore, one may conclude that, for the column-mode imputed data set, the dimensionality of the HONC is likewise one. Referring to Table 4, it can be seen that the factor loadings for the dominant first factor were between 0.436 and 0.930.

4.2.3. Data set with imputation of missing values by row mode value.

The total effective sample size for the row-mode imputed data set was N=502. In like manner to the first two data sets, the Eigenvalue of the first principal component was 6.51, explaining 65.1 % of the total variance, thus lending strong support to unidimensionality of the HONC for this version of the data. The factor loadings for the dominant first factor were between 0.441 and 0.927 (see Table 4).

4.2.4. Data set with imputation of missing values using the EM algorithm.

Finally, almost identical results were obtained for the EM-imputed data set with a total effective sample size of N=501. Note that the EM imputed data set has one less case. The reason for this is that in order to impute data using this algorithm, one must have a minimum amount of information available. That is, if too much information is missing, the EM algorithm cannot produce an estimate of the missing value. Similarly to the above version of the data, the Eigenvalue of the first principal component was 6.52, explaining 65.2 % of the total variance. The factor loadings for the dominant first factor were between 0.435 and 0.929.

To summarize this section, the dimensionality of the HONC was shown to be clearly one for all four versions of the data, that is, irrespective of the missing data technique applied. This finding justifies using a HONC total scale score in the subsequent DIF analyses. Having established the factor structure of the HONC to be unidimensional for all four versions of the data, the next section provides DIF results based on binary logistic regression analysis.

Table 4. Factor loadings of the HONC items for the four versions of the data.

	Without Imputation		**Column Mode Imputation**		**Row Mode Imputation**		**EM Imputation**	
HONC Items	Factor 1	Unique Var	Factor 1	Unique Var	Factor 1	Unique Var	Factor 1	Unique Var
1	.761	.421	.759	.424	.761	.421	.768	.410
2	.839	.295	.816	.333	.821	.325	.832	.308
3	.579	.664	.545	.703	.640	.591	.637	.594
4	.892	.204	.876	.233	.879	.228	.876	.233
5	.912	.169	.908	.175	.914	.165	.910	.172
6	.924	.147	.930	.135	.927	.141	.929	.138
7	.858	.265	.863	.255	.868	.246	.862	.257
8	.676	.543	.658	.566	.656	.570	.656	.569
9	.841	.293	.831	.309	.829	.313	.831	.309
10	.468	.781	.436	.810	.441	.806	.435	.811

Note: The columns titled Factor 1 denote the factor loadings on the first factor.

The columns titled Unique Var denote the amount of error variance.

4.3. Differential Item Functioning (DIF) based on binary logistic regression analysis after matching on the HONC total score

4.3.1. Data set without imputation of missing values.

For the original data set without imputation of missing values, N = 424 cases were left for the analysis. The 2-df Chi-square difference test for DIF (Step 3 minus Step 1) had a significant p-value for item 4 (χ^2 (2, N = 424) = 9.03, p = 0.01, R^2 = 0.015), item 9 (χ^2 (2, N = 424) = 8.45, p = 0.01, R^2 = 0.016), and item 10 (χ^2 (2, N = 424) = 8.66, p = 0.01, R^2 = 0.031). However, according to the effect size criteria suggested by Jodoin and Gierl (2001), the corresponding R^2 values for the 2-df Chi-square difference test for DIF for items 4, 9 and 10 were too small (i.e., below the 0.035 cut-off value for negligible DIF) to flag these items as displaying DIF. For all other items, the 2-df Chi-square difference test for DIF was non-significant. Table 5 shows the χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all HONC items for the data version without imputation of missing values (the three items with a significant p-value are in bold font in Tables 5 to 8).

Table 5. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all HONC items without imputation of missing values.

HONC Item	$\chi^2(2)$	p-value	R^2
1	2.93	0.23	0.006
2	0.08	0.96	0.000
3	0.00	1.00	0.000
4	**9.03**	**0.01**	**0.015**
5	6.09	0.05	0.018
6	2.03	0.36	0.007
7	6.14	0.05	0.009
8	3.41	0.18	0.009
9	**8.45**	**0.01**	**0.016**
10	**8.66**	**0.01**	**0.031**

Tables 9 to 12 provide more details regarding the model-fitting Chi-square values for all three steps (on which the 2-df test of DIF was based), as well as the corresponding Nagelkerke R^2 values obtained for items 4, 9 and 10 at each step for the four versions of the data.

4.3.2. Data set with imputation of missing values by column mode value.

For the data set based on column-wise mode imputation of missing values, (N=502), the results were very similar. As in the above (complete case) analysis, the 2-df test of uniform DIF was significant for item 4 (χ^2 (2, N = 502) = 8.98, p = 0.01, R^2 = 0.013), item 9 (χ^2 (2, N = 502) = 10.94, p = 0.00, R^2 = 0.018), and item 10 (χ^2 (2, N = 501) = 10.34, p = 0.01, R^2 = 0.032). However, according to the effect size criteria suggested by Jodoin and Gierl (2001), the R^2 values for the 2-df Chi-square difference test for DIF for items 4, 9 and 10 are too small to flag these items as displaying DIF. For all other items, the 2-df Chi-square difference test for DIF was non-significant. For comparison, Table 6 shows the χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items with column mode-imputation of missing values.

Table 6. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for column mode-imputation of missing values.

HONC Item	**$\chi^2(2)$**	**p-value**	**R^2 change**
1	2.64	0.27	0.005
2	0.01	1.00	0.000
3	0.05	0.98	0.000
4	**8.98**	**0.01**	**0.013**
5	6.19	0.05	0.007
6	3.83	0.15	0.002
7	3.70	0.16	0.006
8	4.96	0.08	0.011
9	**10.94**	**0.00**	**0.018**
10	**10.37**	**0.01**	**0.032**

4.3.3. Data set with imputation of missing values by row mode value.

The 2-df Chi-square difference test for DIF was again significant for item 4 (χ^2 (2, N = 502) = 8.68, p = 0.01, R^2 = 0.013), item 9 (χ^2 (2, N = 502) = 9.24, p = 0.01, R^2 = 0.016), and item 10 (χ^2 (2, N = 502) = 10.07, p = 0.01, R^2 = 0.029). As with the above versions of the data, the R^2 values for the 2-df Chi-square difference test for DIF for items 4, 9 and 10 are too small to flag these items as displaying DIF. For all other items, the 2-df Chi-square difference test for DIF was again non-significant. Table 7 shows the χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items with row mode-imputation of missing values.

Table 7. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for row mode-imputation of missing values.

HONC Item	**$\chi^2(2)$**	**p-value**	**R^2 change**
1	2.75	0.25	0.006
2	2.75	0.25	0.006
3	0.27	0.88	0.001
4	**8.68**	**0.01**	**0.013**
5	6.25	0.04	0.009
6	4.56	0.10	0.002
7	3.98	0.14	0.018
8	4.49	0.11	0.005
9	**9.24**	**0.01**	**0.016**
10	**10.07**	**0.01**	**0.029**

4.3.4. Data set with imputation of missing values using the EM algorithm.

The 2-df Chi-square difference test for DIF was significant for item 4 (χ^2 (2, N = 501) = 8.79, p = 0.01, R^2 = 0.013), item 9 (χ^2 (2, N = 501) = 10.01, p = 0.01, R^2 = 0.017), and item 10 (χ^2 (2, N = 501) = 10.02, p = 0.01, R^2 = 0.031). Once again, the R^2 values for the 2-df Chi-square difference test for DIF for items 4, 9 and 10 are too small to flag these items as displaying DIF. For all other items, the p-values 2-df Chi-square difference test for DIF was again non-significant. Table 8 shows the χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for EM imputation of missing values.

Table 8. χ^2 values for the 2-df Chi-square difference test for DIF, corresponding p-value and R^2 change values for all ten HONC items for EM imputation of missing values.

HONC Item	**$\chi^2(2)$**	**p-value**	**R^2 change**
1	11.69	0.00	0.020
2	0.10	0.95	0.000
3	0.22	0.90	0.001
4	**8.79**	**0.01**	**0.013**
5	6.37	0.04	0.012
6	3.78	0.15	0.006
7	3.86	0.14	0.009
8	4.66	0.10	0.010
9	**10.01**	**0.01**	**0.017**
10	**10.02**	**0.01**	**0.031**

In summary, across the different methods of missing value imputation (i.e. for all four versions of the data), HONC items 4, 9 and 10 had significant p-values for the 2-df χ^2 test. However, it is emphasized that for these items, even though the 2-df χ^2 tests were significant, the corresponding effect sizes did not reach the cut-off value for DIF, according to the effect size criteria by Jodoin & Gierl, 2001. Thus, these items were not classified as displaying DIF. The next section provides the results of the second set of DIF analyses, based on graphical displays obtained from TestGraf.

Table 9. DIF results of logistic regression analysis without imputation of missing values: Model fitting Chi-square values, df and Nagelkerke R^2 values for items 4, 9 and 10.

Item 4			
Step 1	Chi-square = 252.85	df = 1	$R^2 = 0.60$
Step 2	Chi-square = 261.75	df = 2	$R^2 = 0.62$
Step 3	Chi-square = 261.88	df = 3	$R^2 = 0.62$
Item 9			
Step 1	Chi-square = 205.23	df = 1	$R^2 = 0.52$
Step 2	Chi-square = 213.44	df = 2	$R^2 = 0.53$
Step 3	Chi-square = 213.69	df = 3	$R^2 = 0.53$
Item 10			
Step 1	Chi-square = 36.39	df = 1	$R^2 = 0.14$
Step 2	Chi-square = 37.52	df = 2	$R^2 = 0.14$
Step 3	Chi-square = 45.05	df = 3	$R^2 = 0.17$

Table 10. DIF results of logistic regression analysis with column-wise imputation of missing values: Model fitting Chi-square values and Nagelkerke R^2 values for items 4, 9 and 10.

Item 4			
Step 1	Chi-square = 282.92	df = 1	$R^2 = 0.58$
Step 2	Chi-square = 291.85	df = 2	$R^2 = 0.59$
Step 3	Chi-square = 291.90	df = 3	$R^2 = 0.59$
Item 9			
Step 1	Chi-square = 231.36	df = 1	$R^2 = 0.50$
Step 2	Chi-square = 242.18	df = 2	$R^2 = 0.52$
Step 3	Chi-square = 242.29	df = 3	$R^2 = 0.52$
Item 10			
Step 1	Chi-square = 37.62	df = 1	$R^2 = 0.12$
Step 2	Chi-square = 39.83	df = 2	$R^2 = 0.13$
Step 3	Chi-square = 47.98	df = 3	$R^2 = 0.15$

Table 11. DIF results of logistic regression analysis with row-wise imputation of missing values: Model fitting Chi-square values and Nagelkerke R^2 values for items 4, 9 and 10.

Item 4			
Step 1	Chi-square = 285.89	df = 1	$R^2 = 0.58$
Step 2	Chi-square = 294.46	df = 2	$R^2 = 0.60$
Step 3	Chi-square = 294.57	df = 3	$R^2 = 0.60$
Item 9			
Step 1	Chi-square = 232.61	df = 1	$R^2 = 0.50$
Step 2	Chi-square = 241.84	df = 2	$R^2 = 0.52$
Step 3	Chi-square = 241.85	df = 3	$R^2 = 0.52$
Item 10			
Step 1	Chi-square = 38.77	df = 1	$R^2 = 0.12$
Step 2	Chi-square = 41.28	df = 2	$R^2 = 0.13$
Step 3	Chi-square = 48.85	df = 3	$R^2 = 0.15$

Table 12. DIF results of logistic regression analysis with imputation of missing values using the EM-algorithm: Model fitting Chi-square values and Nagelkerke R^2 values for items 4, 9 and 10.

Item 4			
Step 1	Chi-square = 283.02	df = 1	$R^2 = 0.58$
Step 2	Chi-square = 291.63	df = 2	$R^2 = 0.59$
Step 3	Chi-square = 291.80	df = 3	$R^2 = 0.59$
Item 9			
Step 1	Chi-square = 233.74	df = 1	$R^2 = 0.50$
Step 2	Chi-square = 243.74	df = 2	$R^2 = 0.52$
Step 3	Chi-square = 243.75	df = 3	$R^2 = 0.52$
Item 10			
Step 1	Chi-square = 37.60	df = 1	$R^2 = 0.12$
Step 2	Chi-square = 39.81	df = 2	$R^2 = 0.13$
Step 3	Chi-square = 47.62	df = 3	$R^2 = 0.15$

4.4. Graphical representation of DIF using Nonparametric Item Response Theory (NIRT): DIF assessed with TESTGRAF

4.4.1. Data set without imputation of missing values.

Using the cut-off value of .0415 for composite DIF (based on the guidelines for cut-off indices by Zumbo and Witarsa (2003) discussed in the analysis section), eight out of the 10 HONC items, that is, all but items 2 and 3, were flagged as displaying DIF in this version of the data. Table 13 displays the composite DIF values obtained from TestGraf for each HONC item and for all four versions of the data. Items with composite DIF values in bold font were flagged as displaying DIF. As can be seen from Table 13, DIF was found consistently for the same HONC items, irrespective of the versionof the data (i.e. missing data techniques used). Note, however, that HONC item 6 (see shaded area in Table 13) was flagged as displaying DIF in the data sets without imputation and with row-wise imputation of missing values, but was not flagged as displaying DIF in the versions using column-wise and EM imputation of missing values. Table 14 provides a comparison of binary logistic regression (logR) and TestGraf results (β) across different methods of missing value imputation.

Table 13. TestGraf Composite DIF values for the HONC items for all four versions of the data.

HONC Item	No Imputation	Column Mode Imputation	Row Mode Imputation	EM Imputation
1	**.064**	**.066**	**.065**	**.065**
2	.022	.024	.027	.029
3	.031	.023	.037	.040
4	**.087**	**.089**	**.085**	**.085**
5	**.049**	**.049**	**.049**	**.048**
6	**.043**	.038	**.044**	.039
7	**.046**	**.044**	**.044**	**.043**
8	**.051**	**.063**	**.063**	**.061**
9	**.084**	**.090**	**.084**	**.086**
10	**.074**	**.073**	**.071**	**.066**

Table 14. Comparison of logistic regression (logR) and TestGraf results (β) across different methods of missing value imputation.

HONC Item	**Without Imputation**		**Column Mode Imputation**		**Row Mode Imputation**		**EM Algorithm**	
	Log. R.	**β**	**Log. R.**	**β**	**Log. R.**	**β**	**Log. R.**	**β**
1	_	X	_	X	_	X	_	X
2	_	_	_	_	_	_	_	_
3	_	_	_	_	_	_	_	_
4	0	X	0	X	0	X	0	X
5	_	X	_	X	_	X	_	X
6	_	X	_	_	_	X	_	_
7	_	X	_	X	_	X	_	X
8	_	X	_	X	_	X	_	X
9	0	X	0	X	0	X	0	X
10	0	X	0	X	0	X	0	X

X = Item flagged as displaying DIF according to cut-off indices

- = Item not flagged as displaying DIF

0 = Item flagged for sig. p-value, but DIF effect size below cut-off value for DIF

4.4.2. Data set with imputation of missing values by column mode value.

As can be seen from Table 13, and using the cut-off value of .0415 to classify items as DIF, HONC items 1, 4, 5,7,8,9, and 10 were flagged as DIF. Item 6, which displayed DIF in the previous version of the data without missing value imputation, was now no longer classified as DIF in the version with column-wise imputed missing values.

4.4.3. Data set with imputation of missing values by row mode value.

The results for this version of the data resembled those for the version without imputation of missing values very closely. Except for items 2 and 3, all HONC items fell above the designated cut-off value of .0415 to be flagged as displaying DIF as computed by TestGraf.

4.4.4. Data set with imputation of missing values using the EM algorithm.

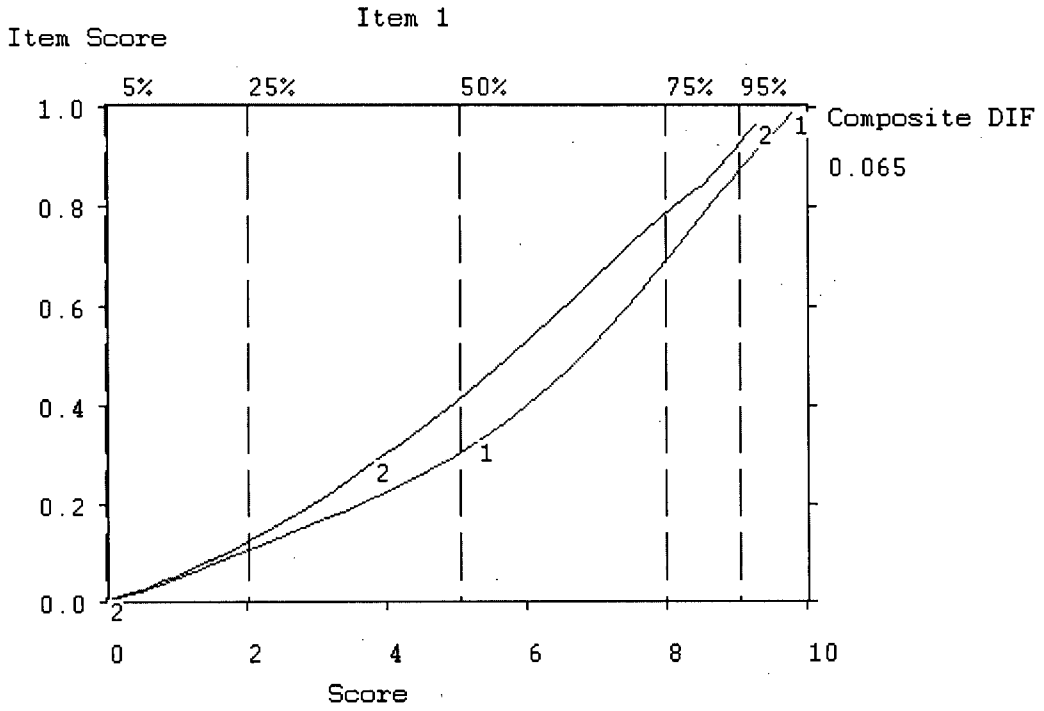
Finally, the results for the version of the data containing EM-imputed estimates of missing values closely resembled the results obtained for the column-mode imputed data version. That is, item 6 no longer was classified as displaying DIF, whereas all other HONC items except items 2 and 3 had composite DIF indices above the cut-off value of .0415.

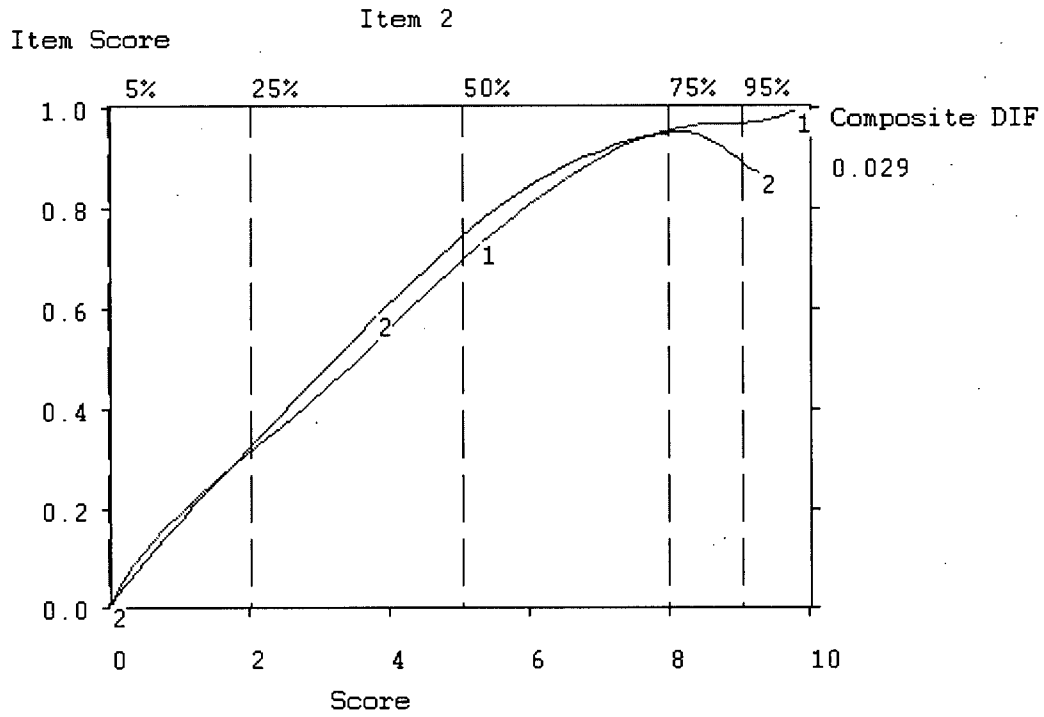
Figure 3 displays the ICCs produced by TestGraf for each of the ten HONC items in the data set with EM imputation of missing values, while Figure 4 depicts the corresponding expected score plot for males and females across the test for the EM imputed version. In like manner, Figure 5 displays the ICCs produced by TestGraf for the ten HONC items in the data set with column-wise imputation of missing values, while Figure 6 depicts the corresponding expected score plot again. In each of the ICC displays provided in Figures 3 and 5, the curve denoted 1 represents the group of female

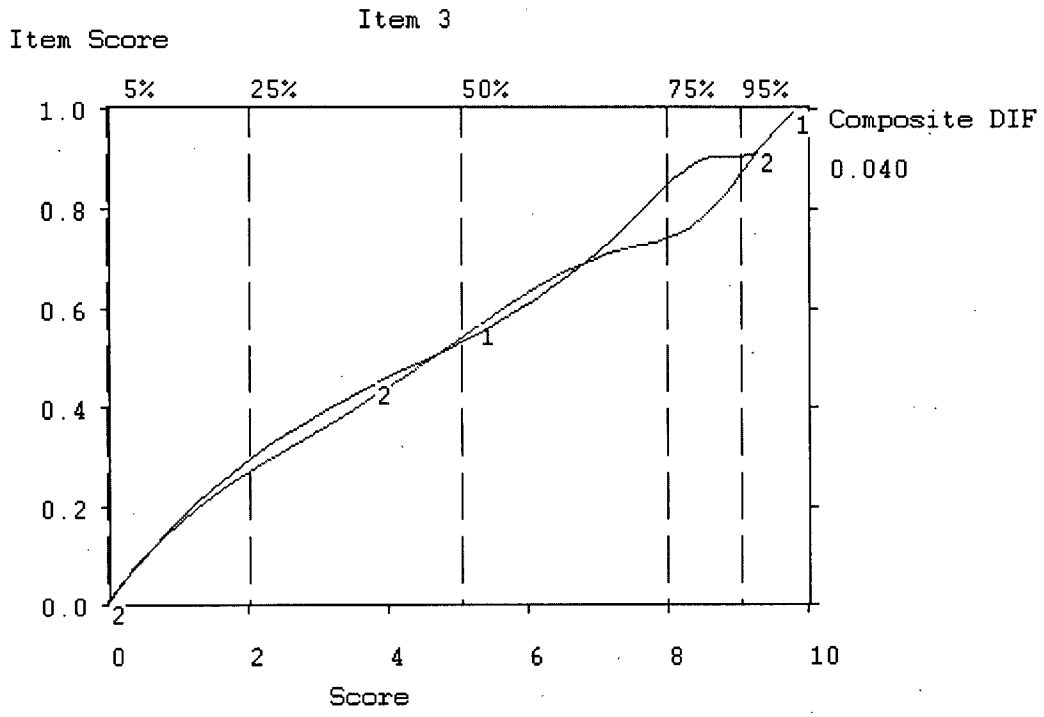
adolescents, while the curve denoted 2 represents the group of male adolescents. Note that, although these two versions of the data were chosen as examples to display all ten ICCs, the results were almost identical with regards to the appearance of the ICCs in the other two versions of the data, that is, using no imputation and using row-wise imputation of missing values.

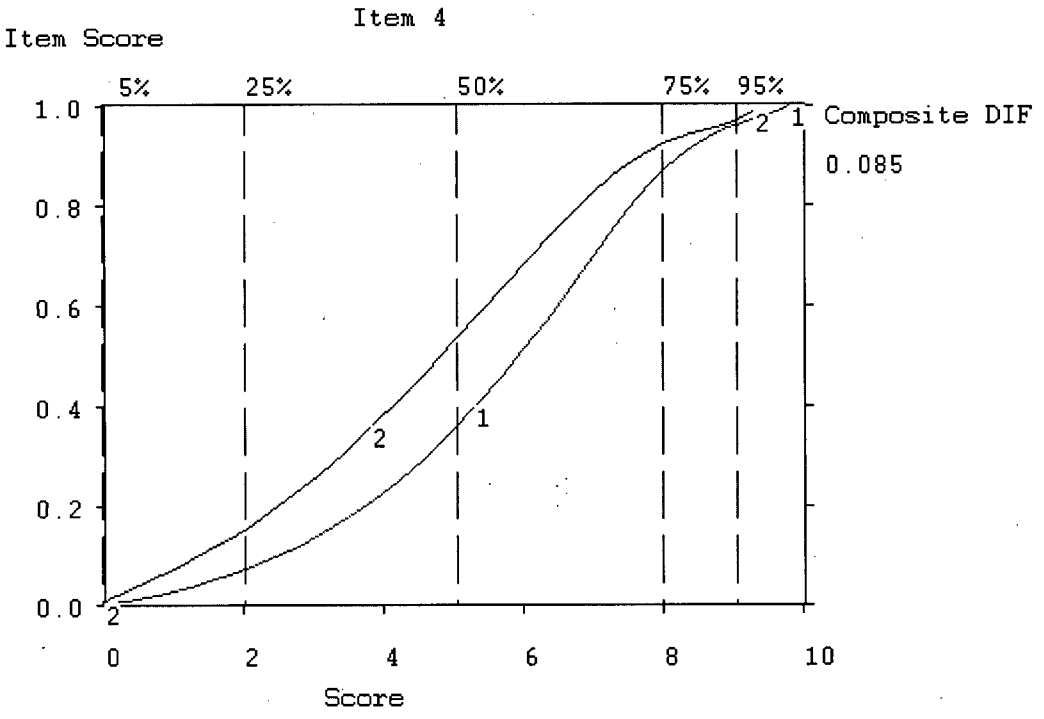
To summarize this section on DIF as assessed by TestGraf, substantially more items (70 to 80 %) were flagged as displaying DIF using this graphical representation of DIF, compared to the binary logistic regression analyses in the previous section, where none of the HONC items was found to display DIF. In the TestGraf analyses, HONC item 6 was also found to be merely on the verge of displaying DIF versus no DIF in the versions with no imputation and row-mode imputation, as the cut-off values showed so little variation and the cut-offs only marginally exceeded the cut-off criterion for DIF as provided by Witarsa and Zumbo (2003).

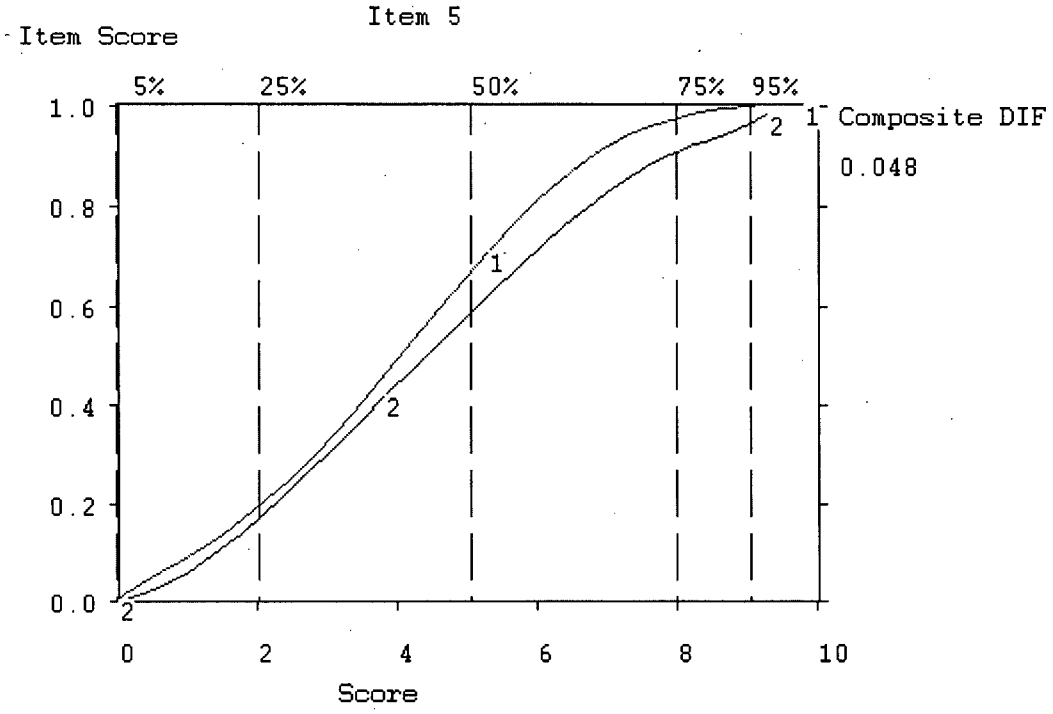
Figure 3. Plots of the TestGraf ICCs of all ten HONC items for the EM imputed data version.

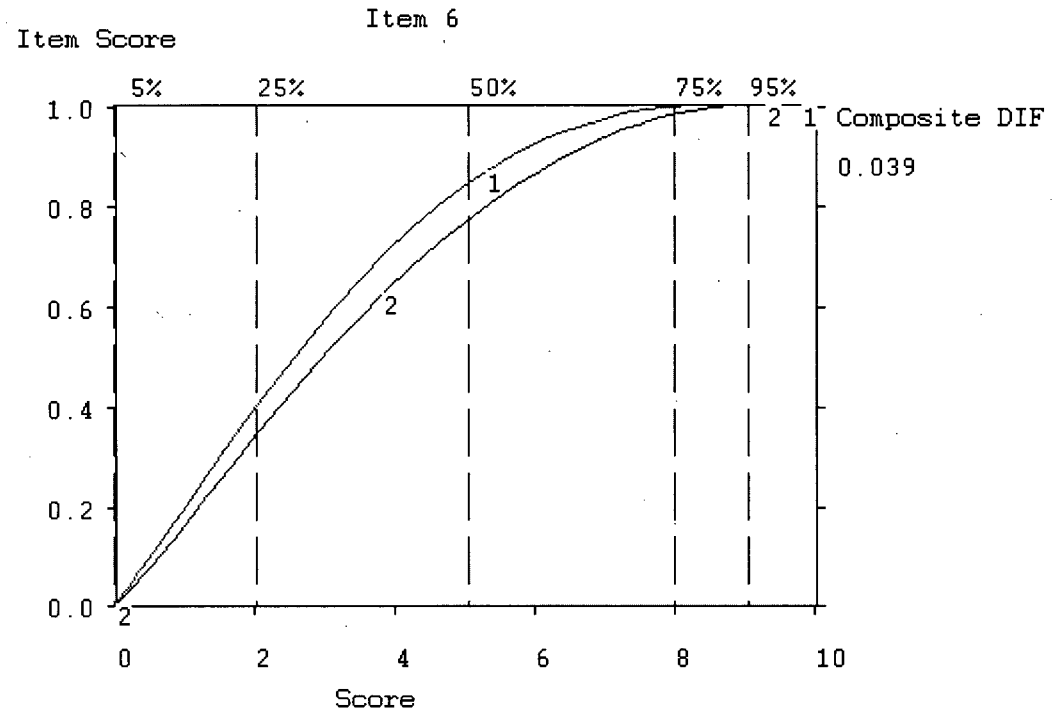


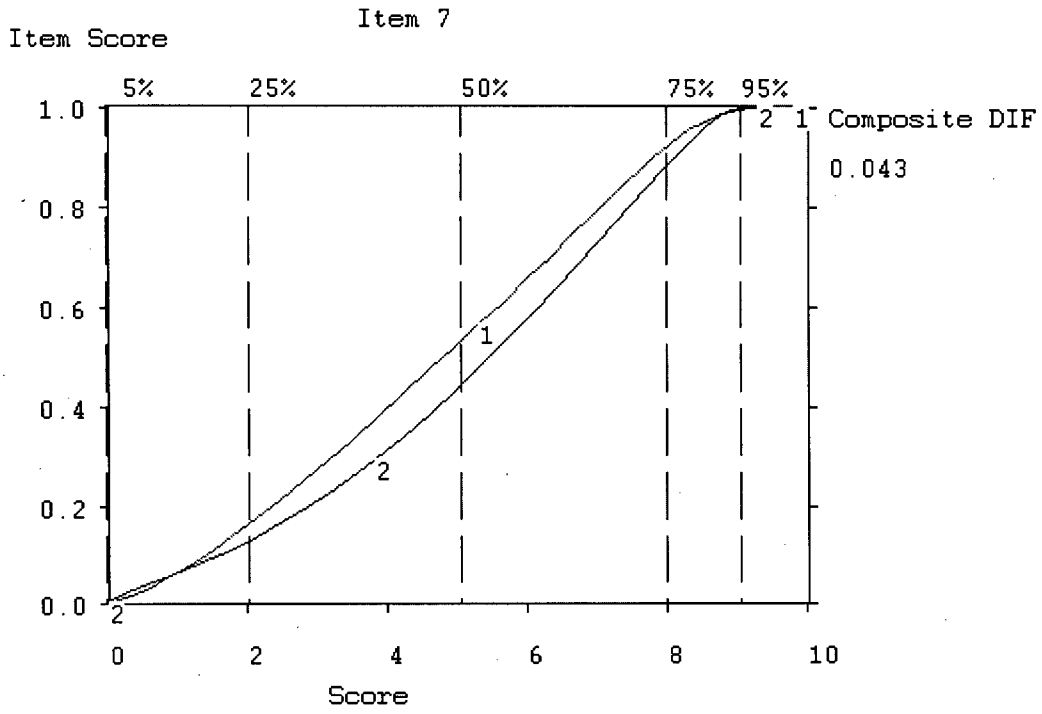


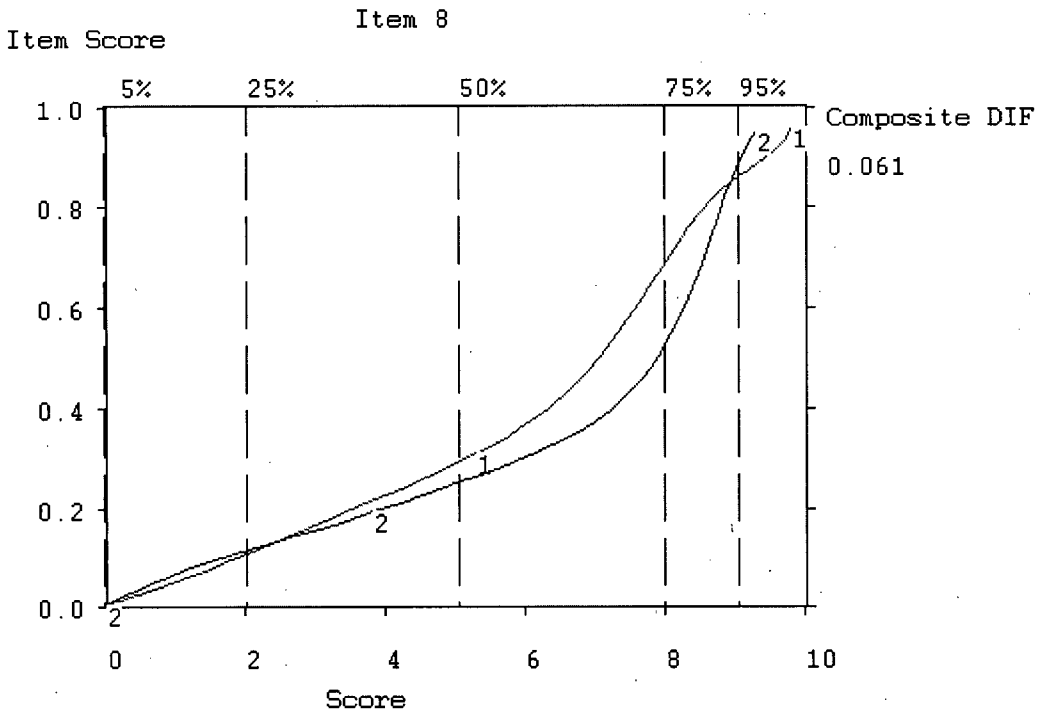

Item 3
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.040
0
2
4
6
8
10
Score

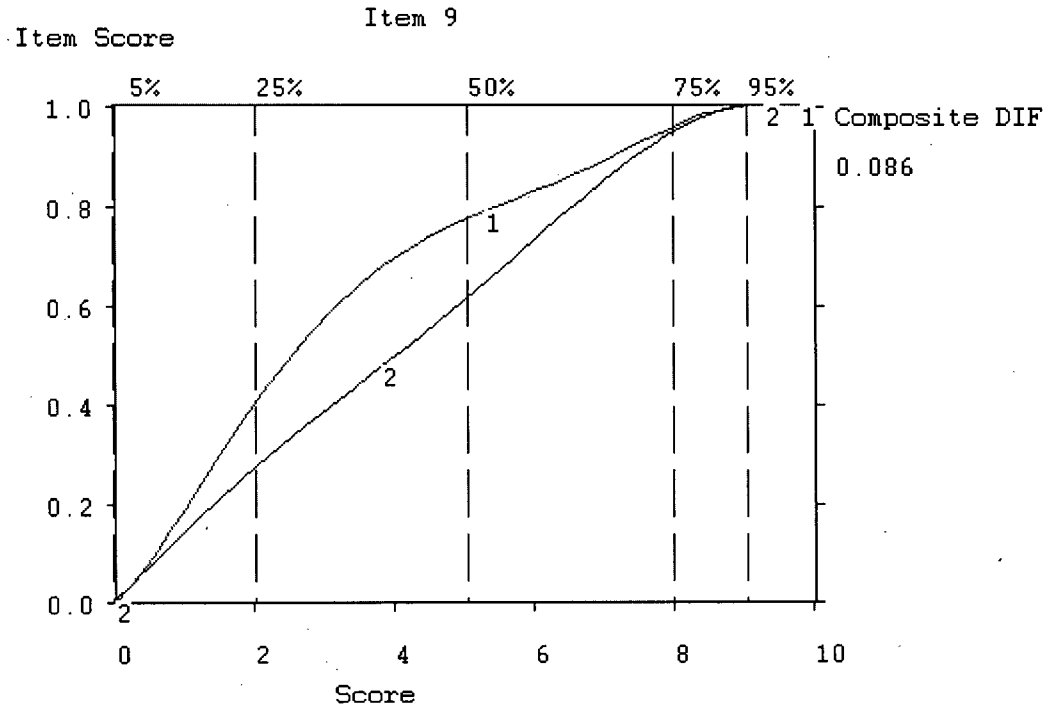

Item 4
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.085
0
2
4
6
8
10
Score


Item 5
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
1
2
Composite DIF
0.048
0
2
4
6
8
10
Score


Item 6
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
2
1
Composite DIF
0.039
0
2
4
6
8
10
Score


Item 7
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
2 1
Composite DIF
0.043
1
2
2
0
2
4
6
8
10
Score


Item 8
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.061
2 1
1
2
2
0
2
4
6
8
10
Score


Item 9
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
2 1
Composite DIF
0.086
1
2
2
0
2
4
6
8
10
Score

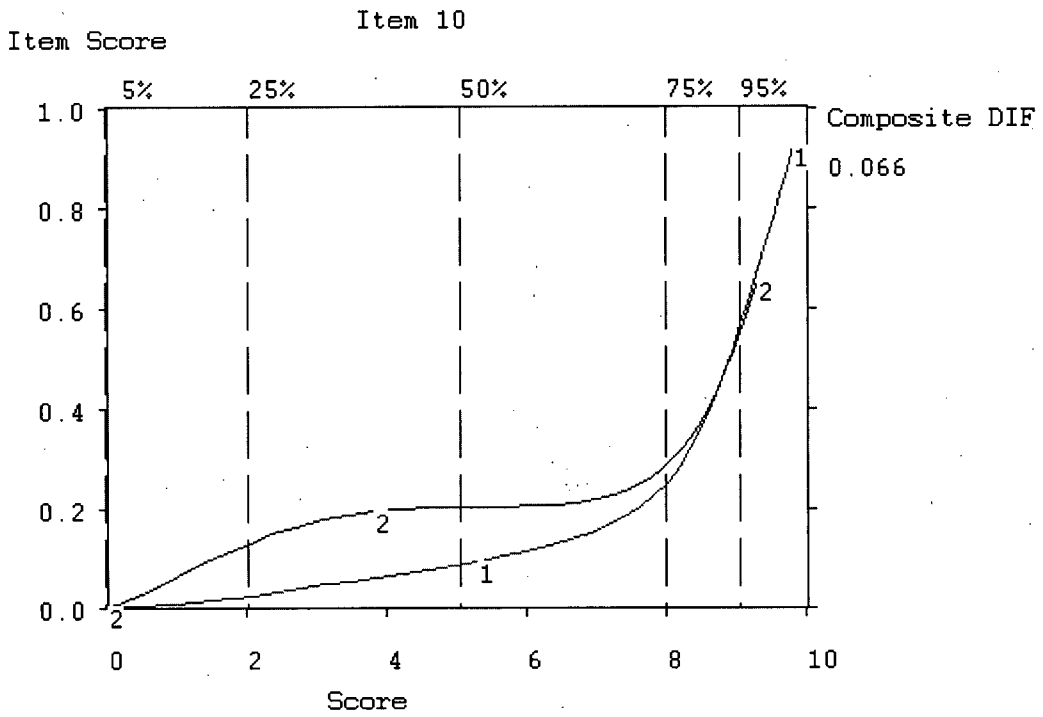
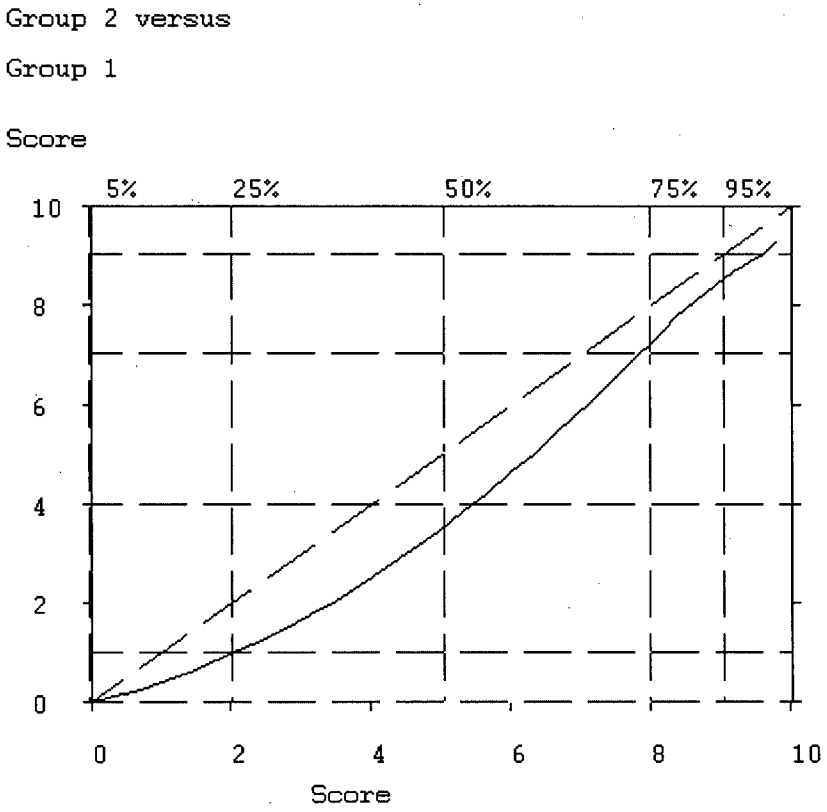

Item 10
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.066
1
2
2
1
2
0
2
4
6
8
10
Score

Figure 4. Expected score plot for the EM imputed version of the data.



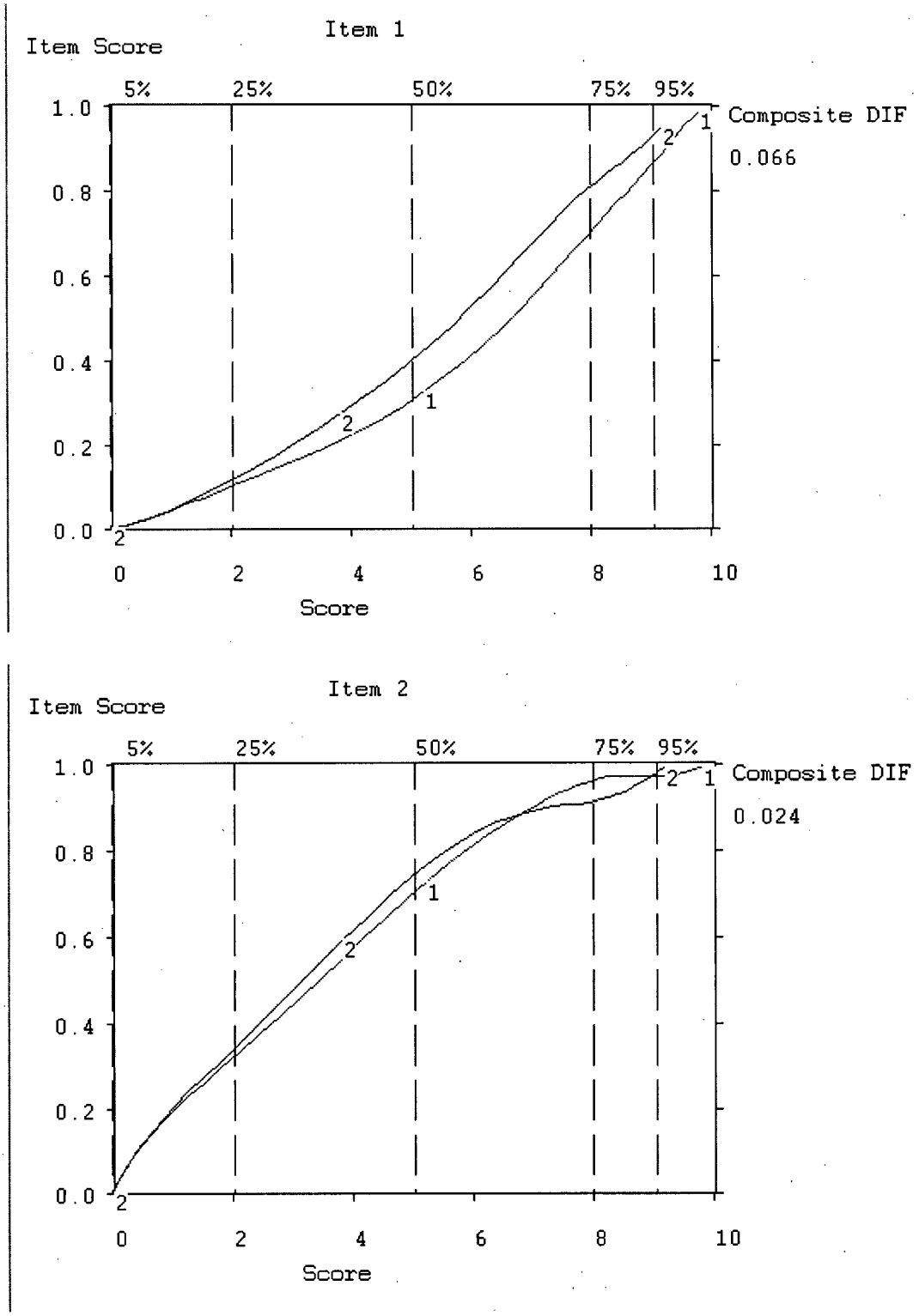
Group1=female which is the x-axis

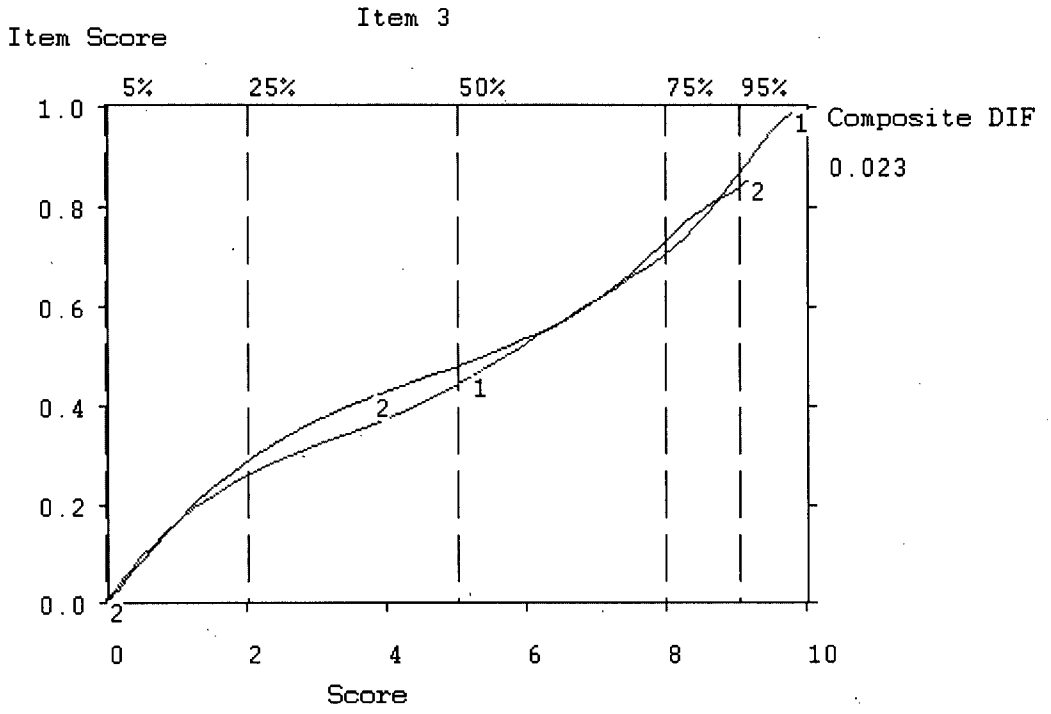
Group 2=male which is the y-axis

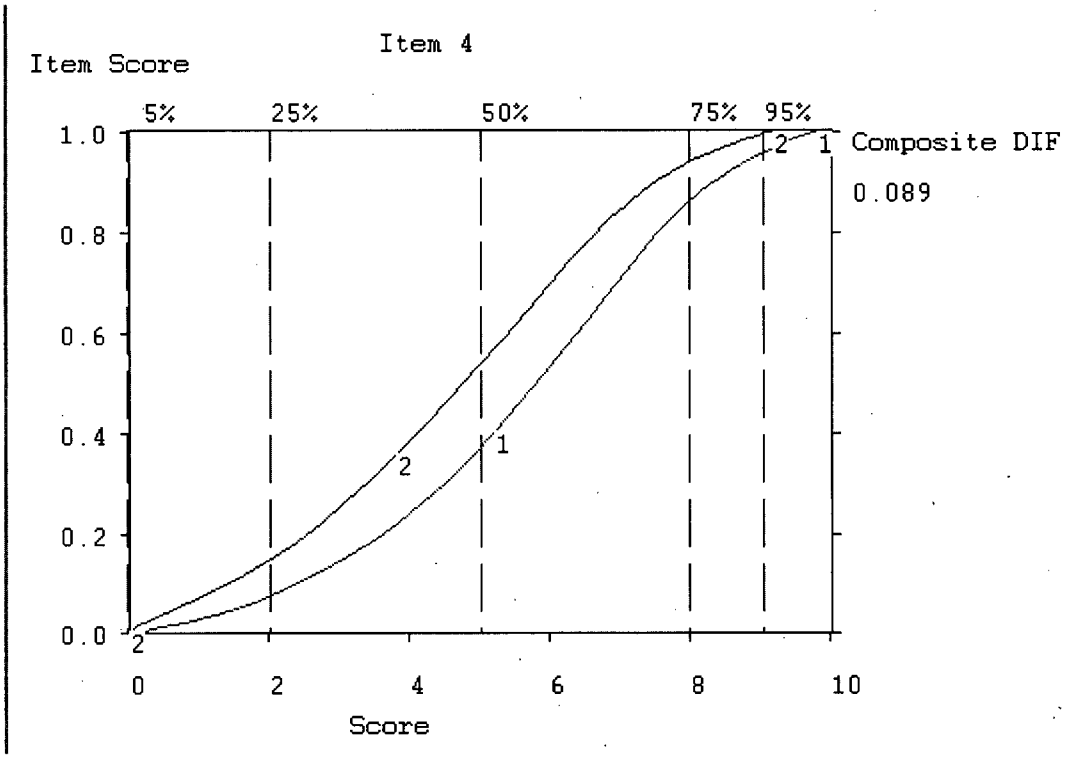
The above plot may be referred to as a plot of expected scores for both groups (males and females) across the whole test. The dashed vertical lines represent the percentile ranks, the x-axis represents the scores for females (group 1), and the y-axis represents the scores for males (group 2). TestGraf produces this plot, which allows one to examine the overall effect of DIF beyond the item level. Thus, the expected score plot is a very informative tool as it provides information as to the overall measure under consideration, and about DIF at the test level. The plot helps convert expected total scores from one group (females) to another (males), thus allowing the linkage of the test as a whole for

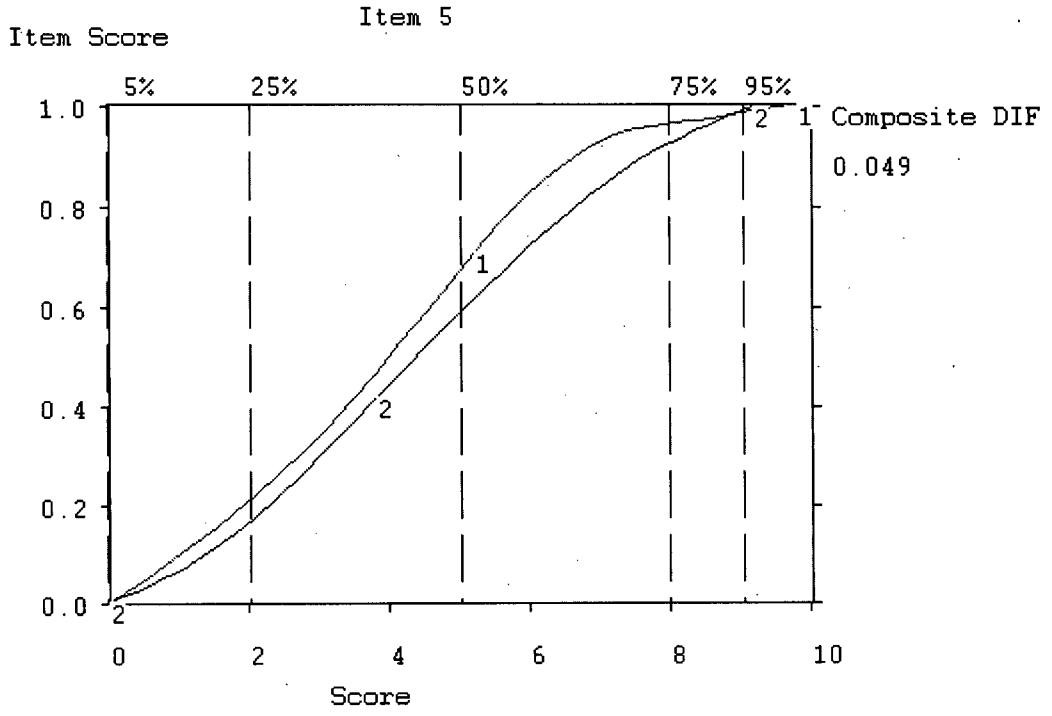
the two groups. Specifically, the above plot can be used to get a sense of the magnitude of DIF at the test level of the HONC by interpreting the percentiles. For example, at the 50th percentile, a score of 5 on the x-axis (females) would correspond to a score of approximately 3.5 on the y-axis, representing males. Likewise, at the 95th percentile, a score of 9 for females would correspond to a score of approximately 8.5 for males. Thus, it can be seen that there is a 0.5 to 1.5 point difference overall in scores obtained on the HONC between females and males, with females scoring higher than males. This finding is consistent with the literature reporting that female adolescents consistently appear to endorse more symptoms of ND than males, as discussed in the literature section of this thesis.

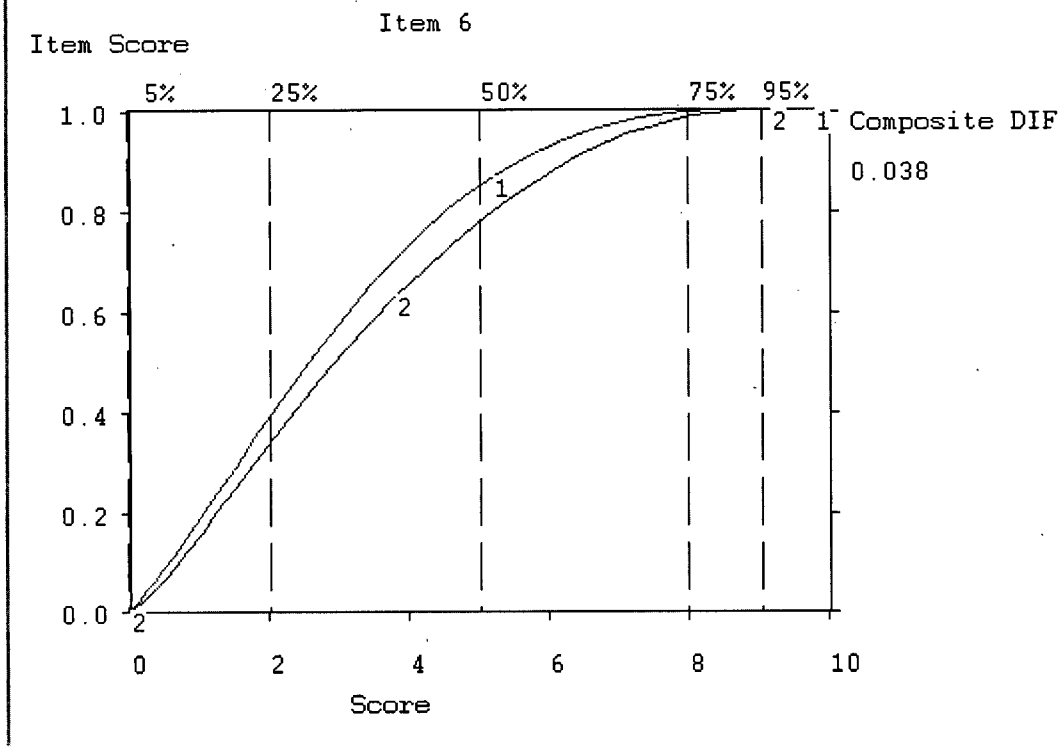
Figure 5. The TestGraf ICCs of all ten HONC items for the data version with column mode imputation.

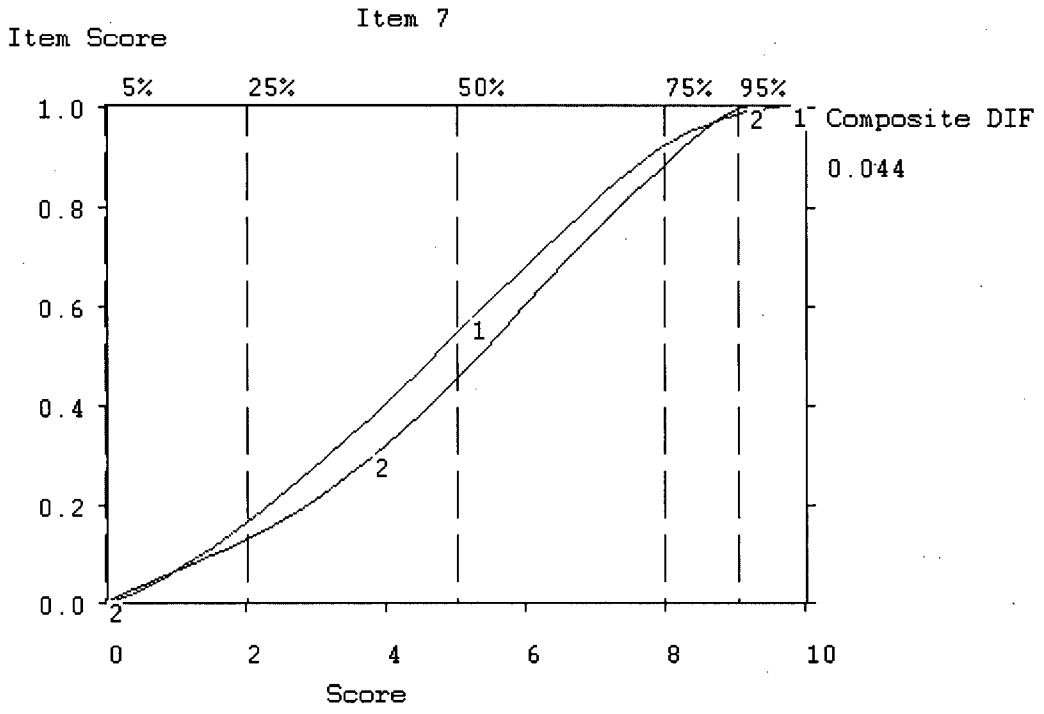


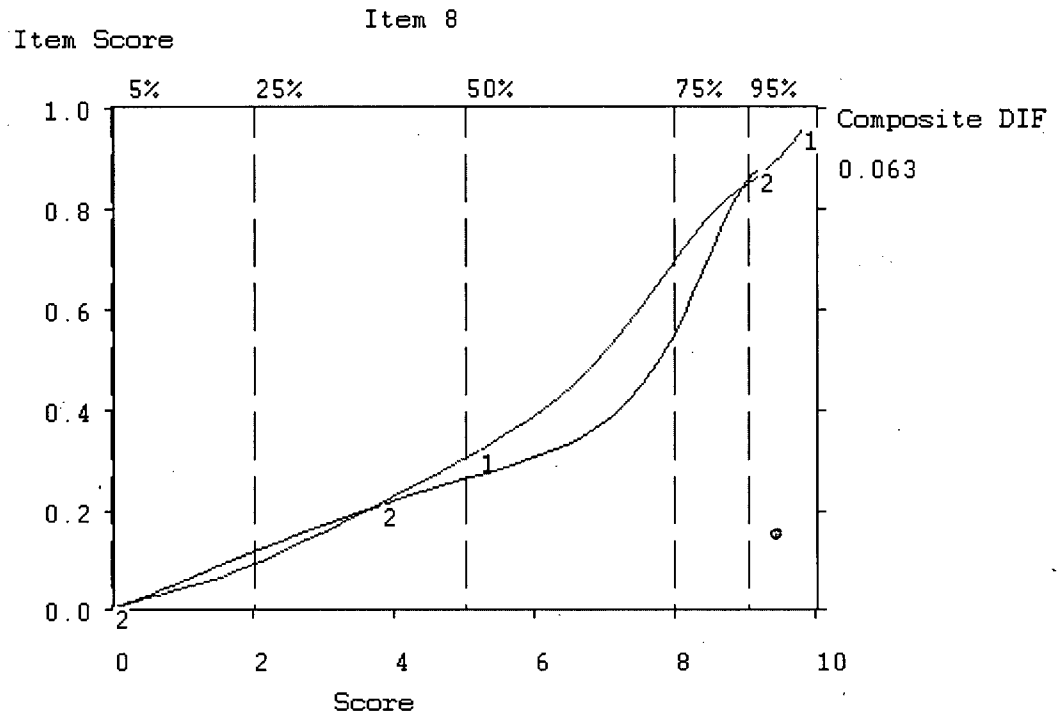

Item 3
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.023
0
2
4
6
8
10
Score

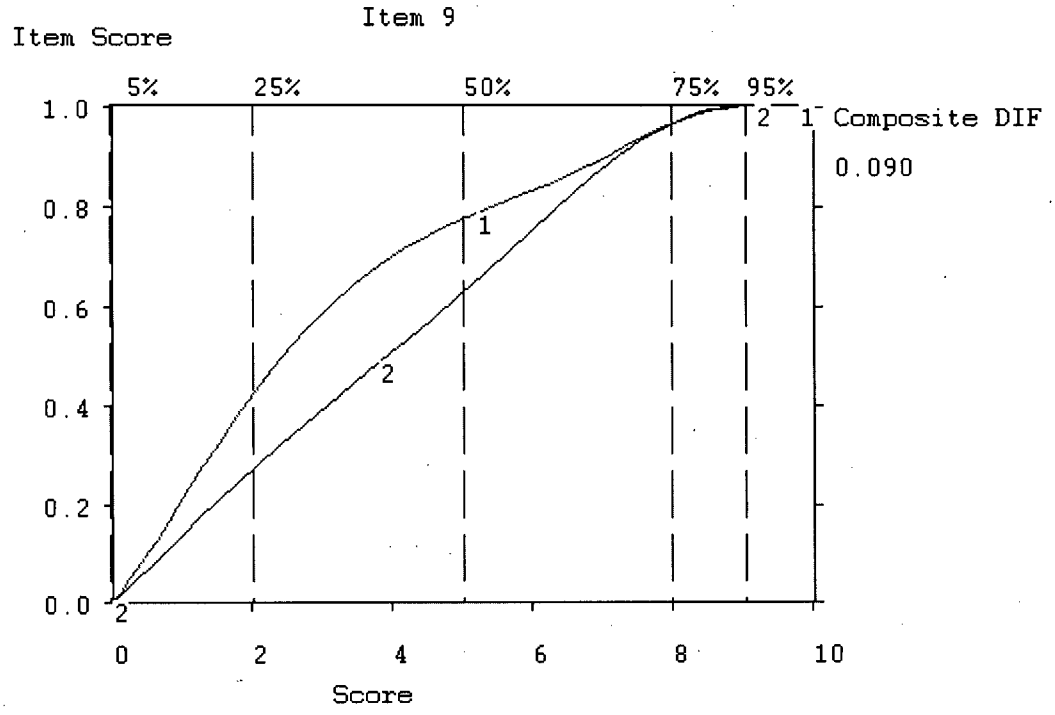

Item 4
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.089
0
2
4
6
8
10
Score


Item 5
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
2
1
Composite DIF
0.049
1
2
2
0
2
4
6
8
10
Score


Item 6
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
2
1
Composite DIF
0.038
1
2
2
0
2
4
6
8
10
Score


Item 7
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.044
0
2
4
6
8
10
Score


Item 8
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.063
0
2
4
6
8
10
Score


Item 9
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.090
0
2
4
6
8
10
Score

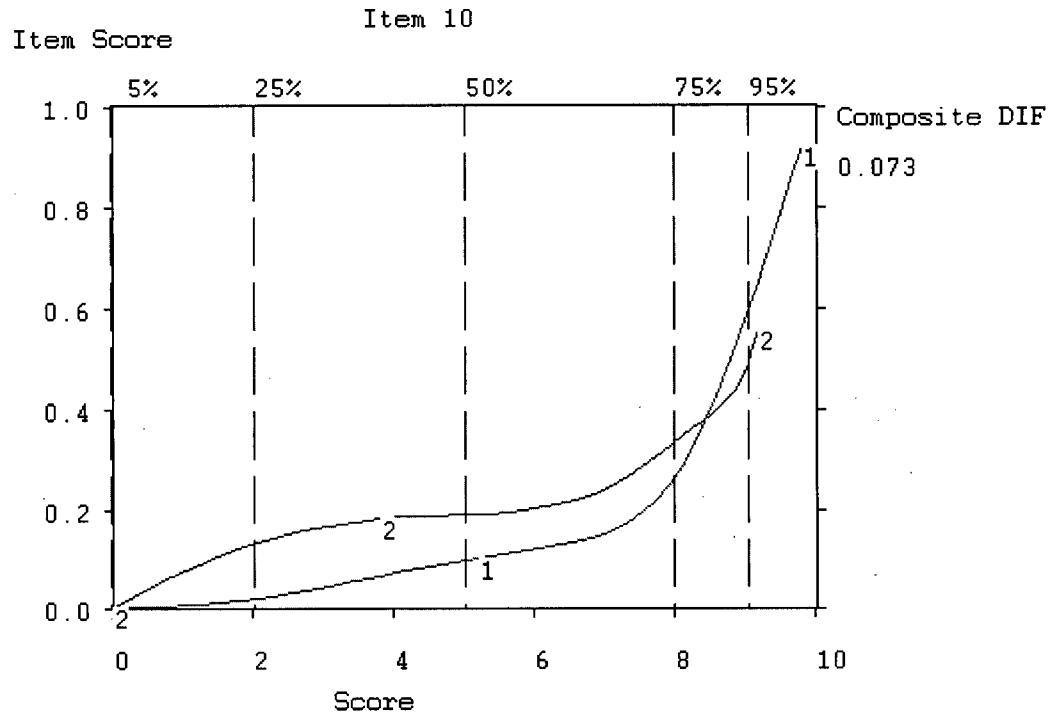
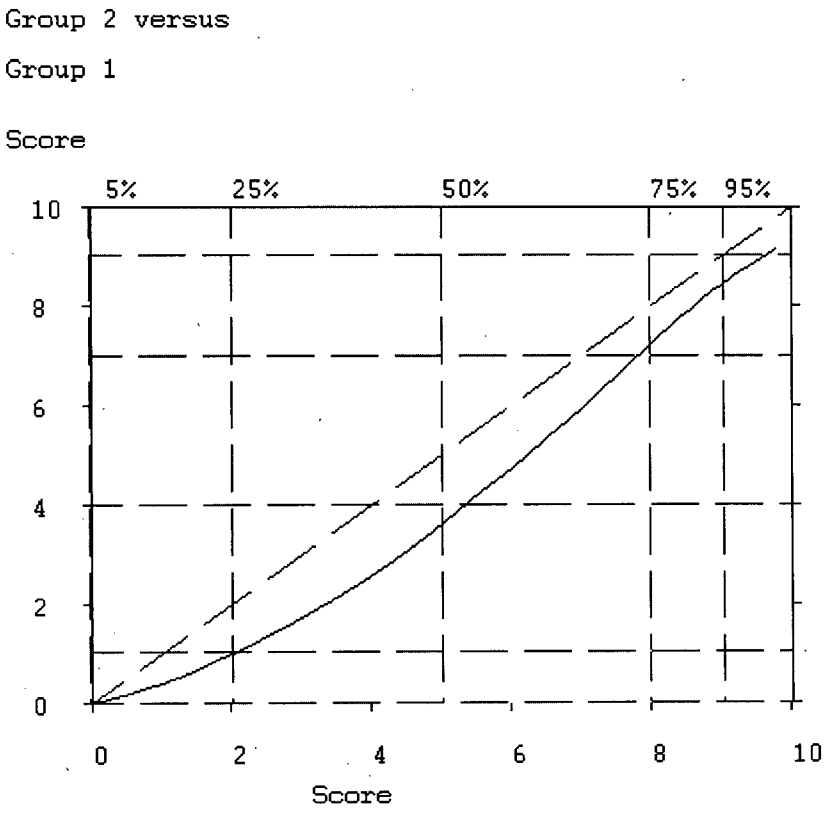

Item 10
Item Score
5%
25%
50%
75%
95%
1.0
0.8
0.6
0.4
0.2
0.0
Composite DIF
0.073
0
2
4
6
8
10
Score

Figure 6. Expected score plot for the column-mode imputed data set.



Group1=female which is the x-axis

Group 2=male which is the y-axis

As with the EM imputed version of the data, the above expected score plot for the column-mode imputed data set provides valuable information about the magnitude of DIF across the HONC at the test level. Again, for example, at the 50th percentile, a score of 5 on the x-axis (females) would correspond to a score of approximately 3.75 on the y-axis, representing males. Likewise, at the 75th percentile, a HONC total score of 8 for females would correspond to a HONC total score of approximately 7 for males. Thus, it can be seen (very much in like manner to the EM imputed example of the data) that there is a 1 to 1.25 point difference overall in total scores obtained on the HONC between

females and males, with females scoring higher than males. This finding is again consistent with the literature on adolescent smoking and ND in general and on the HONC in particular.

Chapter V

Discussion

The main purpose of the present study was to investigate whether gender differences in adolescents' self-reports of ND as assessed by the HONC are a result of measurement artifact due to differential item functioning (DIF). Further, this thesis posed the secondary research question of whether findings of DIF as assessed by two different DIF methodologies are possibly impacted by the various methods for handling missing data used in the analysis.

5.1. Differential item functioning as assessed by statistical modeling using logistic regression

This DIF analysis raised controversial questions about using the cut-off criteria for DIF provided in the literature, and how to go about deciding whether an item may be problematic or not. HONC items 4, 9 and 10 were highlighted in Tables 5 to 8, as these items showed a significant p-value for the 2-df Chi-squared test for DIF. However, according to the effect size criteria used in this study (Jodoin & Gierl, 2001), items 4, 9 and 10 (despite their significant p-value less than or equal to .01) could not be classified as displaying DIF, because their corresponding effect size measure R^2 values were below the cut-off value of 0.035. Nevertheless, it was decided to include these items in the discussion of DIF. Details of the logistic regression DIF analyses for all ten HONC items in all four versions of the data were provided intentionally in Tables 5 to 8, with the purpose of highlighting the size of individual p-values and DIF effect size measure for comparison. Upon inspection of Tables 5 to 8, it becomes immediately apparent that critical consideration needs to be given to how the cut-off indices for DIF are applied.

That is, the cut-off values could be used as strict guidelines, whereby both the 2-df Chi-squared test for DIF and the corresponding measures of effect size represented by the R^2 must meet the criteria for DIF, that is, a p-value of less than or equal to .01 and an R^2 of at least .035. As a consequence, a question of interest to the researcher is whether one can conclude to have no DIF for any of the HONC items--despite a significant 2-df Chi square test for DIF--because the R^2 measure of effect size is too small. If this rule is strictly followed, then (using binary logistic regression), there is no DIF for any of the HONC items in the present study. Applying this rule, this result was found consistently for all four missing data methods applied. In light of these results, however, it is important to keep in mind that a DIF method, such as the logistic regression modeling used here, is but a statistical method for flagging items that are *potentially* biased. Thus, another possibility of using the cut-off criteria for the DIF effect size measure is to apply these suggested cut-off values as *guidelines*, rather than hard-and fast rules for determining the presence versus absence of DIF in a clear-cut manner. Therefore, using the DIF effect size criteria in conjunction with the decision rule of a p-value of less than or equal to .01 as *guidelines* rather than hard-and fast rules, it appears worth highlighting items 4, 9 and 10 (see Tables 5 to 8) as *potentially* problematic, with female adolescents scoring higher than males. It is acknowledged that this is a controversial issue in conducting DIF analyses. However, using the cut-off criteria for the DIF effect size measure as a guideline rather than a strict rule may help researchers identify items for further analysis that may be problematic for one particular group of respondents, rather than concluding that there is no DIF.

Another critical issue to be considered in interpreting the DIF results of the logistic regression, in general, pertains to the assumption that the total score used to match respondents indeed measures what it is intended to measure – the underlying latent variable ND in adolescents. Should this not be the case, covarying out the effect due to ND would not work as intended and thus, would compromise the DIF analyses. It is possible that DIF was not detected to a larger extent due to the matching variable (total) being an imperfect measure of the actual underlying latent variable of interest. It is acknowledged that it is difficult to take a psychometric entity that the items are presumed to capture – ND—and use it as a covariate to be matched upon in the analysis of DIF. However, in any case, this approach is to be preferred to conducting no DIF analyses for the HONC at all.

Finally, in conducting this DIF analysis, it is also important to keep in mind that the groups (female and male adolescents) under investigation were non-randomized groups. Thus, even if clear DIF had been found to a larger extent, one would still not be able to speak in terms of 'causes' of DIF on the HONC items. One would only have identified item(s) as DIF and flagged them for further study. The next section will help emphasize what can be learned by applying more than one method of DIF detection, in addition to using the cut-off criteria for DIF effect size measures as guidelines rather than strict classification rules.

5.2. Differential item functioning as determined by TestGraf

An interesting result was that HONC item 6 was found to be merely on the verge of displaying DIF (row-wise and no imputation) versus no DIF (column and EM imputation), as the cut-off values showed so little variation and the composite DIF values provided by TestGraf only marginally exceeded the cut-off criterion for DIF as provided by Witarsa and Zumbo (2003). Thus, it appears that the method of missing value imputation indeed made a slight difference for the chances of detecting DIF for cases where one has such marginal DIF as displayed by HONC item 6. However, it is worth noting, in this context, that one problem with using the TestGraf program for DIF detection is the lack of a measure of effect size for DIF. What is, in fact, conducted is merely a hypothesis test of significance for DIF.

When comparing the TestGraf results of DIF with the DIF results obtained from binary logistic regression analysis, it is interesting that the latter method did not classify any HONC items as displaying DIF with consistency across all versions of the data. In contrast, TestGraf flagged 7 out of 10 HONC items as displaying DIF for the column-mode and EM imputed data set, and 8 out of 10 HONC items for the row-mode imputed set and the version without missing value imputation. Thus, a researcher using TestGraf as the method of DIF detection for the HONC may come to the conclusion that almost all of the HONC items (i.e., 70 –80%) display gender DIF. In comparison, a researcher using binary logistic regression analyses for DIF detection would have either concluded that there is no DIF (applying cut-off values for the effect size measure as a strict rule) or only marginal DIF for 30% of the HONC items (applying the cut-off values as a guideline). Thus, even though missing data methods did not appear to impact findings of DIF per se,

the *type* of DIF detection method greatly impacted whether an item was flagged as displaying DIF or not.

Finally, it is worth noting that the findings may be sample-specific and cannot be generalized to other samples using the HONC. In the absence of replication data, one cannot be sure about DIF for certain items of the HONC. In order to formulate conclusions about gender DIF present in the HONC, one would have to replicate the results of DIF in another study, or conduct simulation studies.

5.3. Impact of missing data on findings of DIF

The present study did not find major discrepancies in terms of DIF results on HONC items when different missing data methods were applied. That is, DIF results remained largely consistent across the four sets of DIF analyses using various missing data methods. A follow-up question resulting from this finding pertains to the role of missing data and the imputation of missing values. Why did missing data or the method of missing value imputation not have an impact on finding DIF? This question needs to be considered in light of the assumptions as to the pattern of missingness in the data. That is, can a finding of no differences in DIF across different methods of missing value imputation lead to the conclusion that the pattern of missingness is MCAR? This conclusion certainly cannot and should not be drawn, as it cannot be known whether the pattern of missingness is MCAR; there exists no statistical test to determine this. Of what value is the consideration of missingness under the assumptions of MCAR or MAR if, in fact, there exists no statistical method to test these assumptions? The MCAR and MAR assumptions are useful to researchers as mathematical models. However, they are not to be misinterpreted or used as simple practical tools, per se. That is, the MCAR and MAR

assumptions should be used as theoretical models or guidelines for the researcher to critically think about plausible patterns of missingness in his/her data, and to arrive at an informed decision as to how to handle missing data.

Nevertheless, a noteworthy difference between two missing data methods in regards to a finding of DIF became apparent during the TestGraf analyses for HONC item 6 (*When you tried to quit smoking or when you haven't used tobacco in a while, did you or do you feel a strong urge to smoke?).* That is, this item was classified as displaying gender DIF in the versions of the data without imputation and with row-wise imputation of missing values. However, DIF for this item was no longer detected in the data versions with column-wise and EM imputation of missing values (see also Table 13 for these differences). Simulation studies are necessary to clarify the issue of likelihood of detecting DIF when the number of usable cases varies across analyses and using different missing data techniques.

Further, it is not clear why item 3 *(Is it hard to keep from smoking in places where you are not supposed to, i.e. at school?)* had such a large number of missed item responses (12.3% overall), compared to the other HONC items. Interestingly, this item was not identified as displaying gender DIF by logistic regression analysis, and moreover, was one of the only two out of ten HONC items (besides item 2) not displaying DIF in TestGraf graphical displays. Future research should follow up on why item 3 was missed so frequently; subjects could be asked about their understanding of this item.

5.4. Implications: What this study adds

Firstly, the findings from this study have substantiated previously established psychometric properties of the HONC as a measure of ND in adolescents, such as its unidimensionality. However, the present study also corroborated the finding in the literature on the HONC that females, on average, score 1 to 1.5 points higher on the measure of ND than males. This finding is of interest for other researchers using the HONC, as it points out that some items of the HONC may potentially display gender DIF in other studies and populations and may be potentially biased towards females. Further, increased understanding of the construct of ND through DIF analyses has a practical significance from a broader perspective in societal and economic terms. That is, the use of the HONC in research and practice affects decisions for developing effective strategies for smoking prevention and cessation in adolescents. In order to make effective decisions in the practice of smoking cessation and prevention, researchers and policy makers are interested in a psychometrically sound measure of ND, that is, one that is free from item bias.

Finally, the strategy for handling missing data as applied in the present study can be recommended as a sensitivity analysis or as a diagnostic measure to assess the impact of missing values on various sets of analyses. However, it is noted that the present strategy may be less than optimal as a method to impute data. Firstly, the option of complete case analysis often results in few cases usable for the actual analysis. Secondly, the present strategy did not include any multiple imputation of missing values; all versions were based on single value imputation, which is less accurate and applicable to a smaller range of situations than multiple imputation methods.

In light of the results of this psychometric study, what can be recommended to improve the HONC as a measure? Items 4, 9 and 10 should be examined in more detail in follow-up analyses, such as content analyses or qualitative analyses involving adolescents themselves, as these items appear to be potentially problematic in terms of their psychometric performance for males versus females.

Conclusions

This study emphasizes that DIF research for the HONC items is necessary to add new and more detailed psychometric knowledge about this measure of adolescent ND. The results of the present study also contribute to DIF research in general, as the results show that profoundly different findings may be obtained when using different types of DIF methodologies. The use of multiple DIF methods in the present study highlighted the potential difficulties in deciding whether or not an item should be flagged as displaying DIF. The results of DIF for the HONC items in this sample need to be interpreted with caution, despite the consistency of findings when different methods of handling missing values were used. A limitation of the present study was that the two sets of DIF analyses conducted merely helped pinpoint HONC items that are *potentially* gender biased. In the first set of DIF analyses (using binary logistic regression) it could not be ascertained in a clear manner whether or not the HONC items displayed DIF. Further, the impact of missing data on DIF needs to be examined in more depth by conducting simulation studies on the impact of methods for handling missing data on findings of DIF. Finally, given the above mentioned lack of a measure of DIF effect size in the TestGraf DIF detection method, a recommendation for future studies is to work on developing such a

measure of DIF effect size, to be used in conjunction with the graphical representation of DIF in TestGraf.

References

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV, (4th ed.)* Washington, D.C.

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.

DiFranza, J.R., Savageau, J.A., Fletcher, K., Ockene, J.K., Rigotti, N.A., McNeill, A.D., Coleman, M. & Wood, C. (2002a). Measuring the loss of autonomy over nicotine use in adolescents: the DANDY study. *Archives of Pediatric and Adolescent Medicine,* 156, 397-403.

DiFranza, J.R., Savageau, J.A., Fletcher, K., Ockene, J.K., Rigotti, N.A., McNeill, A.D., Coleman, M. & Wood, C. (2002b) Development of symptoms of tobacco dependence in youths: 30 months follow up from the DANDY study. *Tobacco Control,* 11, 228-235.

Ellis, B.B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology,* 74, 6, 912-921.

Enders, C.K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling,* 8(1), *128-141.*

Fagerstrom, K.O. & Schneider, N.G. (1989). Measuring nicotine dependence: a review of the Fagerstrom Tolerance Questionnaire. *J. Behav. Med.,* 12, 159-182.

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education,* 14, 329-349.

Johnson, J., Ratner, P. & Bottorff, J. (2003). Data set containing the Hooked on Nicotine Checklist responses collected within the British Columbia Survey On Smoking and Health.

Johnson, J., Ratner, P.A., Tucker, R. S., Bottorff, J.L., Zumbo, B.D., Prkachin, K.M,. Shoveller, J. (in press). Development of a multidimensional measure of tobacco dependence in adolescence.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87, 1227-1237.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.33-45), Hillsdale, NJ: Erlbaum.

O'Loughlin, J., DiFranza, J., Tarasuk, J., Meshefedjian, G., McMillan-Davey, E., Paradis, G., Tydale, R.F., Clarke, P. & Hanley, J. (2002a). Assessment of nicotine dependence symptoms in adolescents: a comparison of five indicators. *Tobacco Control*, 11, 354-360.

O'Loughlin J. Kishchuk, N., DiFranza, J., Tremblay, M., Paradis, G., (2002b). The hardest thing is the habit: a qualitative investigation of adolescent smoker's experience of nicotine dependence. *Nicotine and Tobacco Research,* 4, 201-209.

O'Loughlin, J., Tarasuk, J., DiFranza, J., Paradis, G. (2002c). Reliability of selected measures of nicotine dependence among adolescents. *Annals of Epidemiology,* 12, 353-362.

Pigott, T.D. (2001). A review of methods for missing data. *Educational Research and Evaluation,* 7, (4), 353-383.

Psujek, J.K., Martz, D.M., Curtin, L., Michael, K.D., Aeschleman, S.R. (2004). Gender differences in the association among nicotine dependence, body image, depression and anxiety within a college population. *Addictive Behaviors*, 29, 375-380.

Rubin, D.B. (1976). Inferences and missing data. *Biometrica,* 63, 581-592.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Normal Data.* New York: Chapman & Hall.

Schillington, A.M., Cotler, L.B., Mager, D.E. & Compton, W.M. (1995). Self-report stability for substance use over 10 years: data from the St. Luis Epidemiological Catchment Study. *Drug and Alcohol Dependence,* 40, 103-109.

Schillington, A.M. & Clapp, J.D. (2000). Self-report stability of adolescent substance use: are there differences for gender, ethnicity and age? *Drug and Alcohol Dependence,* 60, 19-27.

Slocum, Gelin & Zumbo, (2003, in press). Statistical and graphical modeling to investigate differential item functioning for rating scale and Likert item formats.

Swaminathan, H. (1994). Differential item functioning: A discussion. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, and M.W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues.* Ottawa, Canada: University of Ottawa.

Woods, C.M. (2002). Factor analysis of scales composed of binary items: Illustration with the Maudsley Obsessional Compulsive Inventory. *Journal of Psychopathology and Behavioral Assessment,* 24 (4), 215-223.

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., Gelin, M.N. & Hubley, A.M. (2002). The Construction and Use of Psychological Tests and Measures. In the Psychology theme of the *Encyclopedia of Life Support Systems (EOLSS),* Eolss Publishers, Oxford, UK.

Zumbo, B. D., & Hubley, A. M. (2003). Item Bias. In Rocío Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment* (pp. 505-509). Thousand Oaks, CA: Sage Press.

Zumbo, 2004, personal communications, Education Department (Education, Counselling Psychology and Special Education-ECPS), University of British Columbia

Appendix A

Missing Data

Impact of the missing data-problem on statistical analyses and conclusions

A secondary research question in the context of the Hooked On Nicotine Checklist (HONC) was whether a finding of gender DIF is impacted by missing data in the item responses and to what extent. Missing data are a common problem in many social science research contexts (Pigott, 2001). However, the presence of missing data may compromise statistical analyses and conclusions based thereon, as most widely used statistical techniques are not designed for data in which missing values are extensive. If one uses default statistical analyses despite extensive missing data, one risks obtaining misleading results and biased estimates. That is especially the case when one uses default methods provided by computer packages for handling missing data without critically examining the assumptions required of these methods. Specifically, this bias occurs mainly when reasons for the missing data, i.e. missing data mechanisms, are not carefully examined and acknowledged in subsequent analysis steps.

Missing data also make it more difficult to interpret findings, as various response mechanisms can cause missing data in different ways. Thus, all researchers faced with missing data need to give careful consideration to missing data mechanisms and distributional assumptions before proceeding with statistical analyses (Pigott, 2001). For these reasons, with this particular psychometric analysis of the HONC, it is likewise necessary to assess the extent and impact of missing data on the DIF analyses applied in this research.

Prelude to handling missing data: two main issues about the data set at hand

Nature and distribution of the variables of interest.

In general, there are two critical questions that need to be addressed before deciding which missing data imputation techniques can be applied most appropriately to one's data set. The first major question is pertaining to the scale, nature and distribution of the variables in the data set displaying missing data (Pigott, 2001). Is it reasonable to assume that all the variables, including the outcome variables, follow a multivariate normal distribution? The model-based missing data methods in particular require that the data are multivariate normal (Pigott, 2001). This assumption will be discussed in more in the context of maximum likelihood methods discussed in a later section. The next section will emphasize the importance of specifying plausible reasons for the missing values in one's data.

Missing data mechanisms.

The second major issue that needs to be given thought prior to choosing a method for handling missing data is the missing data mechanisms. That is, what are plausible reasons for the missing values? Plausible explanations for missing data serve as important evidence that will help the researcher in making a decision about which missing data method is most suitable for his/her particular analysis. Any method for handling missing data carries specific assumptions about the mechanisms that caused the missing values (Pigott, 2001). For example, the MCAR assumption must be met in order for complete case analysis to be a valid procedure. Complete case analysis is defined as a procedure by which only those cases are included in the analysis that have no missing values in any of the variables of interest (Pigott, 2001). Therefore, complete case analysis excludes all

cases for which there are missing data in any of the variables of interest. This feature is also the main disadvantage of complete case analysis, as it may leave the researcher with too few cases left for the actual analysis. The term missing completely at random (MCAR) is used to describe data for which the responses one has are representative of the originally identified set of cases (Pigott, 2001). That is, MCAR implies that the reasons for missing data are *not* related to any variables (i.e., outcome variables or predictors) in the data set at hand. Under this missing data scenario, complete case analysis is a valid procedure, as the complete cases represent a random sample of the originally identified set. Results from complete cases analyses under the MCAR scenario are thus generalizable to the target population. However, it is stressed again that the main disadvantage of this method is that one obtains less precise estimates than initially aimed for, as one is working with an often much smaller subset of cases for estimation than planned for (Pigott, 2001).

A different scenario applies when the missing data mechanism is not ignorable. In this case, the reasons for missing observations depend on the values of the variables of interest (Pigott, 2002). Under the MAR scenario, complete case analysis would not be an appropriate procedure, as the missing values are not a random sample of the originally identified data set. Missing at random (MAR) describes data that are missing for reasons related to completely observed variables in the data set (Rubin, 1976). The assumption that data are MAR implies that the reasons for missing values on predictor variables are dependent on one or more of the completely observed variables in the data set (Pigott, 2001).

Approaches to handling missing data

The literature on missing data includes several approaches to handling the missing data problem. These procedures may be classified into two broad categories. The first category is briefly outlined in this section. This class of methods includes ad hoc edits, such as single value imputation, complete case analyses (listwise deletion of cases with missing values), and available case analysis (pairwise deletion) (Pigott, 2001). Listwise deletion (LD) results in the use of only complete cases, i.e. cases without missing information. Complete case analysis is chosen either by conscious decision or by default in a statistical analysis, eliminating subjects who have incomplete data on the variables of interest (Pigott, 2001). The problem with using these missing data methods is that they require assumptions about the data that are often violated in practice (Pigott, 2001). That is, complete case analysis (LD) has the main disadvantage that the researcher cannot foresee whether there will be enough cases for analysis based on the number of cases that indeed observe all variables of interest.

Available case analysis or pairwise deletion (PD) uses all available data to obtain estimates of model parameters (Pigott, 2001). In like manner to LD, available case analysis is a valid procedure only when the data are MCAR. Under this scenario (when the remaining data are representative of the originally identified data set), it has been shown that PD provides consistent estimates only when variables are moderately or weakly correlated (Little, 1992). That is, correct point estimates are obtainable using PD under the MCAR scenario. A downfall of PD is that the procedure can produce estimated covariance matrices that are implausible, such as estimating correlations outside of the range of -1.0 to 1.0. These are errors of estimation, occurring due to the differing

numbers of observations entering into estimation of covariance matrix components (Pigott, 2001). A main problem with using PD is that the researcher cannot ascertain or predict when PD will provide adequate results, which is generally viewed as a major downfall of the procedure (Pigott, 2001).

Finally, missing values are often replaced by a plausible value, such as the mean for cases that observe the variable of interest. While this method has the advantage of the inclusion of all cases in the analyses, it also has a major downfall. Specifically, replacing missing values with a single value -the mean for all cases- also changes the distribution of that variable as it decreases the variance (Pigott, 2001). It is acknowledged that imputation of missing data in this manner may yield incorrect imputed values. In addition, when one uses default methods for missing data such as listwise deletion without considering the assumptions required of these methods, one risks obtaining misleading results. For example, the distribution of variables in the data set and the reasons for missing data are two critical issues that need to be considered so that the appropriate missing data techniques can be applied (Pigott, 2001). However, the above-described approaches are conducted to be able to draw comparisons of their impact on DIF results, taking into consideration which method of handling missing data was applied.

It is apparent that the ad hoc methods for handling missing data are not applicable to a wide range of situations and research contexts. Thus, the next section focuses on model-based missing data methods, their underlying assumptions and shortcomings.

Model-based methods for handling missing data

The second broad class of missing data approaches focuses on model-based methods for handling missing data. More specifically, this class of methods focuses on maximum likelihood (ML) methods for dealing with problems caused by missing data, as well as multiple imputations (MI) of missing values, and the expectation maximization (EM) algorithm (Enders, 2001).

In general, one feature of maximum likelihood methods is that they are based on strong distributional assumptions and models for the particular data set at hand (Pigott, 2001). The reason for this is that model-based missing data methods are by necessity based on the multivariate relationship between variables to obtain estimates for the missing values (Pigott, 2001). That is, the model is the multivariate normal distribution, and thus the joint distributions of all variables, including outcomes and predictors, are required to be multivariate normal. At first, this assumption appears to limit the use of these methods for non-ordered categorical variables. However, Schafer (1997) points out that the assumption of multivariate normality can be relaxed to a more flexible assumption that the data are multivariate normal, conditional on the fully observed nominal variables. For categorical variables, methods based on the multivariate normality-assumption should not be used if categorical variables display high rates of missing values (Pigott, 2001). This may be viewed as another limitation of model-based missing data methods, and in certain research situations this is certainly the case. Nevertheless, given that the underlying distributional assumptions hold, model-based methods such as maximum likelihood methods have the great advantage of being appropriate for a wider range of situations than the above discussed ad hoc methods, such as complete case analysis or single value imputation (Pigott, 2001).

Finally, it is important to note that for both maximum likelihood and multiple imputation methods for missing data, the response mechanism is required to be ignorable (Pigott, 2001). That is, model-based missing data methods yield trustworthy results only under the condition that distributional assumptions *and* assumptions for the underlying response mechanism hold (Pigott, 2001). The next section will provide a more detailed description of a maximum likelihood approach, using the EM algorithm.

Handling missing data with maximum likelihood methods using the EM algorithm.

As one of the model-based methods, the EM algorithm naturally requires the assumptions that a) the missing data mechanism is ignorable and b) the joint distribution of the data is multivariate normal (Pigott, 2001). The principle of maximum likelihood must also apply for the case of complete data, when one estimates means or regression coefficients. What is being maximized is the likelihood of the observed data. However, in the case of missing data, the likelihood of observed data is a more complex issue. That is, when missing data occur, it is difficult to maximize the likelihood of the observed data. The researcher is then faced with the problem of obtaining a best estimate that maximizes the likelihood of the observed data (Pigott, 2001). One solution is the EM algorithm, an iterative procedure that finds parameter estimates (e.g. means or covariance matrices) when it is not possible to obtain closed form solutions to a likelihood maximization (Dempster, Laird and Rubin, 1977). This method bases estimation of missing values on the likelihood of the observed data (Pigott, 2001). EM stands for a two step-procedure: E = Expectation, M = Maximization. The EM algorithm thus splits the estimation problem into two estimation steps. The estimation or E-step computes the expected value of the sum of the variables with missing data. It is assumed that one has a value for the

population mean and the variance-covariance matrix. The maximization or M-step uses the expected value of the sum of a variable to estimate the population mean and covariance. Thus, in each step one assumes that one knows one of the two desired pieces of information, that is, either the population mean or the sum of the variable as if it was completely observed. Based on these assumptions, the parameter can be estimated. These steps cycle until the estimates no longer change significantly (Pigott, 2001).

When using maximum likelihood methods such as the EM algorithm, the researcher does not obtain direct values for individual missing variables. What one obtains are estimates for the means and variance-covariance matrix of the variables of interest (Pigott, 2001). Such parameters can then used to obtain model parameters of interest, such as regression coefficients.

In general, when data are MCAR or MAR, the response mechanism is referred to as ignorable (Pigott, 2001). How can one know whether one's data are MCAR or MAR? One cannot obtain direct empirical evidence about the missing data mechanism at work. However, it is possible to examine the sensitivity of results to the MCAR and MAR assumptions by creating several data sets and comparing the analyses thereof. For example, one may run a complete case analysis as a reference point and, in addition, use model-based methods for comparisons. Differences in the results across several analyses may yield important information about the missing data assumptions most relevant to the particular data set (Pigott, 2001).

Appendix B

Factor Analysis of Tetrachoric Correlation Matrices

This appendix will provide a discussion of issues that arise when one requires information on the dimensionality of data that were scored in binary format. In this appendix, the underlying assumptions of factor analysis based on Pearson product moment correlations and the tetrachoric correlation are explained and discussed. Notes of caution regarding the interpretation of results from factor analysis of dichotomously scored variables are provided.

Choosing the appropriate type of factor analysis: A look at the data at hand and the underlying assumptions of factor-analytic techniques

In the following sections, 'linear factor analysis' is defined as factor analysis the way it is typically used to analyze Pearson correlation matrices. Factor analysis based on Pearson product moment correlations is the usual method of choice to obtain proper estimates of the population correlations and information about the underlying dimensionality of continuous data (i.e., variables that have been measured on a continuous scale). However, in the case of dichotomously scored data (Yes/No, T/F, 0/1), one cannot and should not, generally, use linear factor-analytic techniques as designed for use on continuous data. That is, it would not be appropriate to use a matrix of Pearson product moment correlations to obtain information on the dimensionality of one's binary scored (dichotomous) data. Specifically, linear factor analysis models assume that items are linearly related to one another, and that items are linearly associated with the underlying continuous factor. However, both these basic assumptions are likely violated in the case of binary scored items. Further, dichotomously scored items cannot be linearly

associated with continuous factors – another violation of an assumption underlying classical factor analysis when it is used in the case of binary scored variables (Woods, 2002). Finally, the linear factor analysis model assumes that the item responses are normally distributed – a condition that cannot be met under the binary scored scenario, that is, when variables can only take on one of two values (Woods, 2002). Considering the above violations of model assumptions, applying linear factor analysis to dichotomously scored items would be a model misspecification (Woods, 2002).

In general, with binary data there are two classes of solutions (Zumbo, 2004, personal communication). The first solution is based on a linear factor model using a tetrachoric correlation matrix. This approach, fitting the tetrachoric correlation matrix, is the most common and is the analogue of applying ordinary least squares (OLS) regression to continuous data. The second solution is to reproduce the data matrix by choosing a nonlinear model. This solution is the analogue of applying binary logistic regression, but represents an IRT –based method which requires the use of specialized software, such as the program Testfact.

It is clear, then, that in the binary (dichotomously) scored case, an appropriate method of choice is factor analysis of the tetrachoric correlation matrix to obtain information on the dimensionality of the data at hand. A tetrachoric correlation results when the Pearson correlation is applied to dichotomously scored variables (Kubinger, 2003). However, it is assumed that there is a continuous, normally distributed variable underlying the observed binary data. A formula of approximation for the tetrachoric correlation is:

$$r_{tet} = \cos \{180^{o} / [1+\sqrt{(bc/ad)}]\}$$

to be chosen in the case of a, b, c and d so that a and/or d are not zero (Kubinger, 2003). Further, a, b, c and d denote the frequencies in the fourfold contingency table (e.g., a for the counts of cell ++, b for -+, c for +-, and d for --, Kubinger, 2003). The next section describes assumptions underlying the concept of the tetrachoric correlation.

The tetrachoric correlation matrix: Assumptions underlying this type of association between variables

In general, when one uses correlational techniques or factor analysis when dealing with dichotomously scored variables, the main assumption is that these variables in fact represent underlying continua that have been discretized or dichotomized. For example, even a Likert-type scale can consist of as little as two scale points (Rupp, Koh, & Zumbo, 2003). Even though respondents may not feel that a particular item is either completely true or completely false, they still will select "true" if their sentiment towards the item is above a certain threshold along the continuum; otherwise, they will select "false" (Woods, 2002).

Based on the assumption of an underlying continuum, alternative measures of association between variables may be obtained by analyzing the matrix of tetrachoric correlations. For binary scored items, the model is based on Pearson product moment correlations among the items. However, the Pearson coefficients are now called Phi (Φ) coefficients, as both variables to be associated are binary (Woods, 2002). It is important to caution that the Phi coefficients are attenuated (lowered) in the binary scored case, compared to what the correlations would be if the variables were continuous. Specifically, the maximum correlation between two binary items can only be 1.0 for the

(special) case that the items have equal endorsement probabilities (Mislevy, 1986). Factor loadings, in turn, depend on the interim Phi coefficients, and thus, the factor loadings obtained for binary items tend to be underestimated (Woods, 2002). One also needs to consider the fact that, when item endorsement probabilities differ among the items, the linear factor analysis model overestimates the number of factors needed for exploratory-type analyses. This result is due to the fact that items cluster together according to the thresholds along the assumed continuum of variation for the variables, over and above the content measured by the items (Woods, 2002).

The concept of thresholds along an assumed underlying continuum of variation is also important when considering tetrachoric correlations. That is, for a tetrachoric correlation matrix, the assumed underlying continuum for the dichotomous scores represents manifestations of respondents exceeding a certain number of latent thresholds on the underlying continuum (Rupp, Koh, & Zumbo, 2003). One would then estimate the latent thresholds and model the observed cross-classification of response categories (e.g. Yes/No, True/False) using underlying latent continuous variables (Rupp, Koh, & Zumbo, 2003). The next section discusses how one obtains factor loadings under the dichotomously scored scenario.

Factor analytic methods for binary scored items: Types of estimation

The key method for factor analyzing binary scored item responses has been to replace the matrix of Phi coefficients with a matrix of tetrachoric correlations (Woods, 2002). Under this model, tetrachoric correlations are viewed as hypothetical correlations obtained under the assumption that observed item responses represent a truncation of an underlying continuous and normally distributed response process (Cohen & Cohen,

1983). How does one obtain factor loadings under such a scenario? A least squares estimator is used to obtain factor loadings from a tetrachoric correlation matrix (Woods, 2002). There are several types of least squares estimators one can obtain. Unweighted least squares (ULS) estimation is analogous to the ordinary least squares (OLS) estimator used in linear regression. While coefficients are chosen to minimize the squared deviation between observed and predicted values in OLS regression, in ULS factor analysis, factor loadings are chosen to minimize the sum of squared differences between the observed correlation matrix and the matrix of correlations predicted by the model (Woods, 2002). An assumption underlying ULS is, however, that correlations among the items themselves are independent and have constant error variance. For binary items, both of these assumptions are commonly violated. To overcome dependences among correlations and heterogeneous error variances among the tetrachoric correlations, the ULS estimator may be weighted, rendering the estimation procedure "weighted least squares" (WLS) (Woods, 2002). The weight matrix so obtained contains variances of and covariances among the correlations, which corrects for heterogeneous error variances and dependencies among correlations (Woods, 2002).

The following example compares a Pearson correlation matrix with a tetrachoric correlation matrix, based on 493 cases. The items are taken from the Hooked on Nicotine Checklist (HONC) analyzed for this thesis. The first table displays a Pearson correlation matrix; the second table displays a tetrachoric correlation matrix.

Endorsement frequencies for HONC items 1 and 2:

HONC1	Frequency	Percentage
0	319	64.7
1	174	35.3

HONC2	Frequency	Percentage
0	211	42.8
1	282	57.2

Table 1. A matrix of Pearson correlations between HONC items 1 and 2.

	Item 1	Item 2
Item 1	1.00	.51
Item 2	.51	1.00

Table 2. A matrix of tetrachoric correlations between HONC items 1 and 2.

	Item 1	Item 2
Item 1	1.00	.78
Item 2	.78	1.00

Challenges commonly encountered when using binary scored variables in a factor analysis

Even though the latent variable distribution is not necessarily required to be normal, there are limitations to be considered when factor analyzing binary scored variables. That is, one needs to take into consideration to which degree the analysis approach is still reasonable, in that inferential results may not always be invariant or even comparable when the assumption of an underlying normal distribution is violated (Rupp, Koh, & Zumbo, 2003).

Another point is worth considering on theoretical grounds before choosing a factor analytic method when one has only dichotomously scored variables. That is, in some social science research contexts, it is difficult to establish a threshold along the assumed continuum of variation below which respondents would endorse the "No" option, as opposed to the "Yes" option above this hypothetical cut point along the continuum. For example, in order for the technique to be valid, one has to assume that a certain amount of the underlying construct of interest will indeed lead respondents to rate themselves as "addicted", while another amount of latent construct will decisively lead respondents to rate themselves as "not addicted". It is acknowledged that there may be theoretical and conceptual problems when cutting an assumed underlying continuous variable distribution in two halves, each now presumably denoting a discrete category.

Appendix C

Differential Item Functioning (DIF)

Definition of Differential Item Functioning (DIF) and DIF frameworks

DIF is present when a test item functions differentially for one group of examinees (e.g. females) than for another (e.g., males). That is, examinees from different groups show differing probabilities of endorsing an item after matching on the underlying construct that the items intend to measure (Zumbo, 1999). Thus, DIF means that certain groups of respondents endorse items differently with respect to characteristics other than those due to actual differences in the construct being measured. Another way of defining DIF is in terms of measurement invariance, as the test item displaying DIF does not perform the same way for different groups of examinees (Zumbo & Hubley, 2003).

Related to DIF is the concept of item bias. Item bias is defined as examinees from one group having a greater chance of answering an item correctly (or endorsing the item) than examinees from another group due to some characteristic of the test item or situation irrelevant to the testing purpose (Zumbo, 1999). As such, DIF is a required, but insufficient condition for item bias. Thus, if there is no DIF, then there is no item bias. If DIF is present, however, one still may not declare that the item is biased. Instead, a follow-up analysis is needed to empirically determine the occurrence of item bias (Zumbo, 1999). That is, the item(s) flagged for DIF would be submitted to content analysis conducted by content specialists in order to determine whether the test items carry bias towards a particular group of examinees. Based on the content analysis, one would then either revise the item so that it no longer carries bias, or one would choose to eliminate the item(s).

Finally, item impact is the converse concept of item bias, in that item impact is present when examinees from different groups have different chances of endorsing the item due to true differences between the groups in the underlying ability being measured by the item (Zumbo, 1999).

DIF frameworks in the literature.

Several broad frameworks for conceptualizing DIF exist in the literature. The first framework is concerned with statistical modeling of item responses using contingency tables or regression models (Zumbo & Hubley, 2003). As DIF is displayed when persons from one group answer items correctly more often than persons from another group, it follows from this definition that it is necessary to match respondents on the underlying ability of interest before one studies group effects. Therefore, DIF exists when, after conditioning on the differences in item responding *due to the underlying ability being measured*, the two groups (e.g. males and females) still show differences. The regression framework for DIF detection thus aims at stating a probability model for studying main effects ('uniform DIF') and the interaction of group-by-ability ('non-uniform DIF'), after statistically controlling for the total score on the test (Zumbo & Hubley, 2003). An advantage of the regression framework is that the effects of both the grouping variable and the interaction term can be studied simultaneously using conditional methods, that is, conditioning on the total test score in order to determine these effects over and above the total score. This class of methods uses logistic regression models for each individual item. One tests the statistical effects of the grouping variable(s) and the interaction of the grouping variable with the total score after conditioning on the total score.

The second major framework is concerned with item response theory (IRT) models for detecting DIF. In this framework, two item characteristic curves (ICCs) are examined for one item, but computed from two different groups (Zumbo & Hubley, 2003). Within the IRT framework, if the item displays DIF, then the two ICCs will appear different for the two groups.

There are several ways in which the ICCs can differ. First, the two curves can differ in terms of the item difficulty (threshold) parameter (b). If this is the case, the two curves are displaced by a shift in their location along the continuum of variation (i.e., theta). Alternatively, the ICCs can differ on item discrimination (a). If this is the case, the two ICCs will intersect. In this general IRT framework of DIF, the first scenario would represent uniform DIF, whereas the second scenario would represent non-uniform DIF, showing the interaction of group-by-ability (Zumbo & Hubley, 2003).

In general, the IRT approach to DIF focuses on the area between the curves, comparing the IRT parameters of the two groups. This comparison of IRT parameters for groups is an unconditional analysis, as it is assumed that the ability distribution has been 'integrated out' by computing the area between the curves across the distribution of the continuum of variation (Zumbo & Hubley, 2003). Specific IRT methods for detecting DIF are discussed in more detail in the Appendix Nonparametric Item Response Theory (NIRT).

Uses and goals of DIF analyses

In general, DIF methods are used in developing new tests, adapting existing measures and for validating inferences from test scores (Zumbo & Hubley, 2003). Specifically, there are several goals behind DIF analyses (Zumbo, 2004, personal communication).

The first goal of investigating DIF is concerned with fairness and equity in testing, where groups are defined ahead of time by policy and legislation. The second goal is the investigations of DIF to make group comparisons and rule out measurement artifact as an explanation for the group differences. Groups are identified ahead of time, which is often determined by the research question of interest. The main purpose of this type of DIF analyses is to deal with the potential threat to internal validity when items do not function the same for two groups of examinees. The third goal of DIF analyses is to enhance understanding of the cognitive and psychosocial processes underlying item responding. In this case, the main purpose is to determine whether such processes are identical for different groups of respondents.

Impact of DIF on statistical analyses and psychometric issues resulting from DIF

If there is DIF, then there may be item bias. That is, DIF is a necessary but not sufficient condition for item bias. Item bias occurs when examinees of one group are less likely to endorse an item than examinees from another group, due to some characteristic of the test item(s) or the testing situation that is irrelevant to the test purpose (Zumbo, 1999). Thus, DIF is a potential threat to the validity of inferences made from scores, as without validation, inferences made from measures or tests are meaningless (Zumbo, 1999). If DIF is apparent, then one needs to apply follow-up item analysis (e.g., content analysis) to determine the presence of item bias.

Finally, Zumbo & Hubley (2003) warn that, as DIF methods in general are applied in non-experimental or quasi experimental studies, it is important to use caution not to interpret findings of DIF in causal terms, that is, by stating causal claims of grouping variable effects when the study is observational.

Appendix D

Nonparametric IRT

This appendix will briefly describe item response modeling using nonparametric item response theory (NIRT). This approach may be best presented by briefly describing the general framework of parametric IRT before extending it to the broader, more expanded approach of NIRT. In doing so, this appendix is primarily based on Zumbo, Gelin and Hubley (2002). A definition of NIRT as a modeling approach will be presented. Major theoretical and practical motivations for its development and use will be described, and finally, the advantages of applying NIRT will be discussed.

From item response modeling in general to nonparametric item response modeling: What is nonparametric item response theory (NIRT)?

IRT methods, in general, are based on two fundamental assumptions. The first is the notion of a latent variable that is distributed along a continuum of variation. This continuum of variation may be envisioned as a continuum of a quantitative latent variable along which individuals vary. The latent variable is depicted on the continuum of variation, theta (θ). Thus, item response modeling techniques, in general, allow the modeling of item responses as a function of this continuum of variation (Zumbo, Gelin & Hubley, 2002).

The second fundamental assumption of IRT methods, in general, is that an item response function (IRF) links the latent variable score with a probability of responding a certain way (Hambleton, Swaminathan & Rogers, 1991). That is, the person's item response can be linked to the level of the latent variable present; the item response is determined by this quantitative amount. The IRF may thus be conceptualized as the

regression of the item response variable on the latent continuum of variation. Thus, IRT methods, in general, state that the performance of an examinee can be predicted or explained from various factors, such as ability (Hambleton, Swaminathan & Rogers, 1991). The IRF (also called Item Characteristic Curve-ICC) depicts the relationship between the likely item response and the levels of the continuum of variation (Zumbo, Gelin & Hubley, 2002). Via the IRF, it is thus possible to obtain information about persons at the item level over the range of the continuum of variation. The next section focuses on the main features of parametric item response theory (PIRT), to be contrasted with the main features of nonparametric item response theory (NIRT) in a later section.

In parametric item response theory, (PIRT), three important assumptions are made about the nature of the IRF:

1. The dimensionality of the latent space is one; one latent variable can account for the joint item distributions observed.
2. Local independence; this means that, for the case in which we have *k* dichotomous items, the conditional joint probability is the product of *k* conditional univariate probabilities (Zumbo, 2003, Sijtsma, 1998). Local independence also means that, if the test common factor is partialed out from any two items, their residual covariance is zero, therefore implying the product rule stated above.
3. The assumption of a *monotone* increasing IRF; this means that the item responses provided by subjects over all items are ordered by the latent variable(s). That is, the IRFs $P_j(\theta) = P(X_j = 1\text{'given' } \theta)$ are *nondecreasing* as a function of θ.

To summarize, IRT approaches, in general, aim to establish models that account for the likelihood of endorsement for items as a function of the latent variable θ and the

item's characteristics, such as the ability to discriminate among respondents (Zumbo, Gelin & Hubley, 2002). However, very large sample sizes and a large number of items are required to perform PIRT modeling. This is a practical limitation for many research contexts in the social and health sciences. Therefore, NIRT modeling approaches were desired requiring fewer items and subjects, while posing less restrictive assumptions about the nature of the IRF. The next section focuses on motivations for the development of NIRT.

Motivations for the development of NIRT

Overall, there have been three broad motivations for developing NIRT (Junker & Sijtsma, 2001):

1. To delineate a commonality among models of both PIRT and NIRT, their features need to be characterized. These features are local independence (LI), monotonicity of item response functions (IRFs) and unidimensionality of the latent variable. Through the characterization of these model features, it should be discovered what happens when IRT models satisfy only imperfect or weak versions of these features (Junker & Sijtsma, 2001). Further, one can characterize successful and unsuccessful inferences under broad model features in order to formulate conclusions about how IRT models use information from the data and aggregate it. NIRT may be employed to accomplish these goals.

2. A parametric IRT model fit to data is likely to be incorrect. That is, when one has applied a family of PIRT models and it is suspected (or shown) that they fit the data at hand poorly, a more flexible family of models – NIRT models – is desired. NIRT models may be employed to assess violations of LI due to nuisance traits (latent variable multidimensionality) and to identify sources and effects of differential item functioning

(DIF). Further, NIRT models provide a more flexible context for developing methodologies that establish the most appropriate number of latent dimensions underlying a test. Finally, NIRT provides alternatives for PIRT models in tests of goodness of fit.

3. IRT models applicable to smaller sample sizes, such as in psychosocial and sociological research, were desired. NIRT models make more economic use of the data at hand than PIRT models by identifying items that scale together well (i.e., follow a particular set of NIRT assumptions). With NIRT, several subscales with simple structure may be identified among scales in the case where items do not form a single unidimensional scale (Junker & Sijtsma, 2001). This feature is also one of the major advantages that NIRT provides over PIRT.

How is NIRT different from PIRT?

NIRT is fundamentally different from PIRT in that NIRT is akin to nonparametric simple regression, focusing on the individual data points available. In NIRT, the form of the IRF relating the item response y and θ is determined by the data, utilizing the existing data optimally, whereas in PIRT, the IRF is a pre-determined function of a model, such as a logistic function or normal ogive (Zumbo, Gelin & Hubley, 2002). As such, NIRT is a very data-driven technique.

For example, NIRT based on Ramsey's (2000) approach uses a class of nonparametric regression methods that partitions the continuum of variation into intervals, within which the likelihood of an item response may be estimated. Here again, the X-axis represents scores along the latent variable θ, and the Y-axis depicts the likely item response at a given score (level) of θ. This particular type of nonparametric item response modeling

represents a graphical approach by employing graphical displays of nonparametric IRFs. However, in contrast to PIRT, nonparametric item response modeling does not employ parameters such as the intercept (α) and slope (β) in a formal regression equation stating the conditional distribution of the item responses (y).

Resulting from the fact that NIRT is a very data driven approach, that is, following the data points available more closely, NIRT is also fundamentally different from PIRT in terms of the form of the underlying IRF. That is, the PIRT approach is based on an IRF with a specified parametric form. By contrast, the nonparametric IRF is allowed to take any form. The curve is allowed to follow the data points very closely, as opposed to being specified to a parametric shape. Therefore, the nonparametric IRF could be non-monotone increasing or decreasing at certain levels of θ. TESTGRAF, a computer program developed by Ramsay (2000), is used to create the graphical displays of the nonparametric IRFs. In this graphical approach, one is primarily interested in the area between the two nonparametric IRFs representing two groups (e.g., males and females) for a particular item. Further, one may focus on the intervals that break the IRF into sections representing confidence limits at particular points of the continuum of variation (Zumbo, Gelin & Hubley, 2002). These intervals provide information about the likelihood of endorsing the item at different levels of θ, the construct of interest (e.g., tobacco dependence). For example, if the expected scale score on the latent variable ND is zero (respondents do not possess this symptom of ND), then the most likely item response to this questions should be 0 (does not endorse this symptom of ND) and not 1 (does endorse this symptom of ND).

As such, graphical nonparametric item response modeling is not based on numerical values of observed total or aggregate scores in creating the graphical displays. Rather, Ramsay's method using TESTGRAF replaces the observed aggregate scores with their respective ranks. These ranks are then replaced by their corresponding standard normal quantiles (i.e., z-scores). Next, a line smoothing technique called Gaussian Kernel Smoothing is applied to the obtained nonparametric regression line to create the graphical display of the nonparametric IRF. The X-axis of the graphical display may represent the standard normal quantiles or expected scores, such as represented by the original scale.

To summarize, the nonparametric item response modeling approach is different from the general PIRT framework in several respects. Most importantly, it allows for more flexibility as it does not require large sample sizes and item pools as PIRT does. The next section will highlight advantages of using this NIRT framework for analyzing a smaller set of items.

Advantages of NIRT applications over PIRT

As already pointed out at the end of the previous section, the applicability of IRT methods to smaller sample sizes and smaller sets of test items is greatly enhanced through NIRT approaches. For example, in a data set such as the one under investigation, the sample size of 256 (males) and 257 (females) per group and the number of items i = 1...10 would preclude the application of IRT methodology. However, we desire information at various points of the continuum of variation. That is, we are interested in the question of how well the instrument works at different levels of tobacco dependence. This question can only be answered on the item-by-item level. NIRT can provide this information, even for small sample sizes and item pools as mentioned above, by creating

the nonparametric IRF for each item. In relation to this property, one may apply NIRT first, in order to obtain a glance of what the IRF looks like. Based on this graphical information, one may decide on which further item analyses – NIRT or PIRT -- are most suitable. As NIRT is based on fewer assumptions than PIRT, the nonparametric IRF for each item reflect the nature of the data well, that is, the IRF can take a non-monotone increasing or a decreasing form. This information is desired in order to make judgments about the performance of the particular item at various levels of θ.

In addition, it is possible to obtain a conditional reliability estimate at different levels of the construct, (e.g., ND) as opposed to a simple overall reliability based on a total score. This is a desirable property as it provides information on how precisely the items measure the construct and how this measurement precision changes across various levels of the latent variable (Zumbo, Gelin & Hubley, 2002). Therefore, through nonparametric item response modeling techniques for small sample sizes, one obtains more information about item performance in relation to the aggregate score (scale score) than one could obtain applying classical test theory approaches (CTT).