Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 12, April 2015 ISSN 1531-7714

An Introduction to Missing Data in the Context of Differential Item Functioning

Kathleen Banks, *Middle Tennessee State University*

This article introduces practitioners and researchers to the topic of missing data in the context of differential item functioning (DIF), reviews the current literature on the issue, discusses implications of the review, and offers suggestions for future research. A total of nine studies were reviewed. All of these studies determined what effect particular missing data techniques would have on the results of certain DIF detection procedures under various conditions. The most important finding of this review involved the use of zero imputation as a missing data technique. The review shows that zero imputation can lead to inflated Type I errors, especially in cases where the examinees ability level has not been taken into consideration.

The purpose of this article was to introduce practitioners and researchers to the topic of missing data in the context of differential item functioning (DIF), review the current literature on the issue, discuss implications of the review, and offer suggestions for future research. DIF occurs when two or more distinct groups with equal ability differ in their probabilities of answering test items correctly (Holland & Wainer, 1993). Test takers produce missing data on educational assessments by omitting or not reaching one or more of the items. An omit happens when a test taker accidently skips an item, or after reading it, fails to respond to the item. Given that the individual responds to subsequent items, omitted responses occur earlier in a test. A test taker may not reach an item because of lack of time. Since the individual does not respond to subsequent items, not-reached responses occur at the end of a timed test (Ludlow & O'Leary, 1999).

There is a certain paradox to the task of conducting DIF analyses that has not been seriously considered in the educational measurement literature. The very groups for whom DIF analyses are conducted in the interest of (e.g., females) are also the groups who have a tendency to omit or not reach items (Ludlow & O'Leary, 1999). These focal examinees may omit or not reach one or more of the studied items whose responses are needed to determine whether such items function differentially against them in favor of reference examinees (e.g., males). And yet commonly used DIF procedures such as Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and simultaneous item bias test or SIBTEST (Shealy & Stout, 1993) were not designed to handle missing data.

So, how does one approach the task of conducting DIF analyses in the presence of missing data? Person 1 may exclude individuals with any missing item responses from all DIF analyses, leaving only complete cases (listwise deletion). Person 2 may only eliminate subjects from a DIF analysis if they do not respond to the studied item (analysiswise deletion). Person 3 may retain examinees for all DIF analyses by scoring their missing item responses as incorrect (zero imputation). One's choice of missing data technique could become a source of bias; therefore masking true DIF or creating false DIF. For example, the process of excluding individuals with missing item responses could lead to a great reduction in sample size and limit the power to detect DIF if it really exists. Likewise, the process of retaining examinees by scoring their missing item responses as incorrect could result in situations where items show DIF when no DIF is actually present.

The purpose of this article then was to introduce practitioners and researchers to the topic of missing data in the context of DIF, review the current literature on the issue, discuss implications of the review, and offer suggestions for future research. The article is organized as follows. First, the concept of missing data mechanism is presented. Second, common missing data techniques are discussed. Third, common DIF detection procedures are discussed. Fourth, a review of the missing data DIF research is offered. Fifth, implications of the review as well as suggestions for future research are provided.

Missing Data Mechanisms

Rubin (1976) described three probabilistic explanations for why data are missing. These include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are MCAR if there is no justifiable reason for why it is missing. Randomness explains the missing data (Peng & Zhu, 2008). A test could have MCAR data if both focal and reference examinees accidently skipped some items. Data are MAR if the chance of it being missing is systematically related to data that has been observed (Peng & Zhu, 2008). For example, in a 30-item test where Item 1 is the studied item, examinees missing response to Item 1 could be attributed to their group membership (focal, reference), and/or their observed performance on Items 2-29. Data are MNAR if the likelihood of it being missing is systematically related to data that has not been observed (Peng & Zhu, 2008). Using the previous illustration, examinees missing response to Item 1 could be attributed to their potential performance on Item 1 (answering Item 1 right or wrong).

Common Missing Data Techniques

Some common traditional missing data techniques include listwise deletion, analysiswise deletion, zero imputation, and regression imputation. The first three techniques were discussed earlier. Regression imputation retains individuals for all statistical analyses by predicting their missing data values from a linear regression equation that is constructed from observed variables in the dataset (Graham, 2009). In terms of DIF, all subjects are included in every DIF analysis because their missing item responses are predicted from observed variables (e.g., group membership, performance on other items) in the dataset. One common modern missing data technique involves multiple imputation via regression imputation (Graham, 2009). In terms of DIF, each person's missing item response is predicted using existing values from observed data. The process is repeated (maybe 3-10 times) to generate a collection of similar but different plausible values for the missing item response. DIF is then calculated separately on each of the *m* complete datasets to obtain *m* parameter estimates of the amount of DIF present in the studied item. The *m* parameter estimates are averaged across the *m* complete datasets to obtain a single best estimate of the amount of DIF present in the studied item.

Common DIF Detection Procedures

Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993) are the most commonly used procedures for assessing DIF with dichotomous (e.g., multiple-choice) data. Extensive research has been done on these methods and indicated that they require much smaller sample sizes for sufficient parameter estimates than item response theory approaches. Reasonable statistical power and Type I error rates have been observed using complete datasets. All of these procedures can detect uniform DIF in complete datasets, while logistic regression can also detect non-uniform DIF in complete datasets (Clauser & Mazor, 1998). Uniform DIF occurs when the magnitude of DIF against the focal group remains constant along the trait scale. Non-uniform DIF occurs when the magnitude of DIF against the focal group changes along the trait scale. In the case of non-uniform crossing DIF, the focal group is advantaged by the studied item at lower trait levels, while the reference group is advantaged by the same item at higher trait levels (Li & Stout, 1996). Li and Stout (1996) stated, "Although there is the widespread assumption that crossing DIF is relatively rare in practice, many researchers have documented the presence of crossing DIF in real test data (e.g., Bennett, Rock & Kaplan, 1987; Ellis, 1989; Hambleton & Rogers, 1989; Mellenbergh, 1982)" (p. 649). Despite their findings, the DIF literature has focused primarily on uniform DIF at the expense of non-uniform DIF. In fact, out of the nine studies included in this review, only one examined missing data in a non-uniform crossing DIF context, while the remaining studies examined missing data in a uniform DIF context.

Review of Missing Data DIF Research

Peer-reviewed research on the topic of missing data in the context of DIF is scant. The author is aware of four refereed journal articles (Emenogu, Falenchuk, & Childs, 2010; Finch, 2011a, 2011b; Robitzsch & Rupp, 2009), eight national conference presentations (Banks & Walker, 2006; Falenchuk & Emenogu, 2006; Falenchuk & Herbert, 2009; Herbert, Falenchuk, & Childs, 2008; Rousseau, Bertrand, & Boiteau, 2004, 2006; Rupp, Choi, & Ferne, 2006; Sedivy, Zhang, & Traxel, 2006), and two doctoral dissertations (Emenogu, 2006; Garrett, 2009) on the subject. While Emenogu (2006), Emenogu et al. (2010) Study 1, Rousseau et al. (2004), and Rupp et al. (2006) were real-data illustrations, the other studies were simulation experiments. Only the simulation experiments were reviewed. Also, since Emenogu et al. (2010) Study 2, Falenchuk and Emenogu (2006), and Herbert et al. (2008) were duplicates of the same simulation experiment, only the peer-reviewed Emenogu et al. (2010) Study 2 was reviewed. All of these studies determined what effect particular missing data techniques would have on the results of certain DIF detection procedures under various conditions. The nine studies were reviewed according to three characteristics: (1) data generation and examinee conditions, (2) data manipulation conditions, as well as, (3) analysis of conditions and recommendations for practitioners and researchers who are faced with missing data when conducting DIF analyses.

Data Generation and Examinee Conditions

According to Table 1, out of the nine studies, seven generated data by fitting, a 1-, 2-, or 3-parameter logistic item response theory model to dichotomous data and had test lengths between 20 and 40 items. Two studies generated data by fitting a graded response or partial credit model to polytomous (constructed-response) data and had test lengths of 9 or 20 items. Reasonably large sample sizes were created in all of the studies (between 100 and 4,000 examinees per group), with some developing equal and/or unequal sample sizes. It is not uncommon for reference and focal groups to differ in their underlying ability distributions (known as impact) which can have an effect on DIF detection with or without missing data (Finch, 2011b). In five studies, reference and focal examinees were drawn from standard normal distributions (mean = 0, standard deviation = 1) to represent no impact or from normal distributions with different means to represent impact. Three studies simulated no impact as described above and one study did not give any information about the presence or absence of impact.

Data Manipulation Conditions

Examinees produced missing data by omitting or not-reaching one or more of the studied items. Table 2 indicates that eight studies allowed test-takers to omit between one and twenty-five items; however, only one item was studied at a time. One study allowed individuals to omit as described above or not-reach the last set of items on the test. In this case, the not-reached responses were studied together. Two of the studies that investigated omitted responses also took into consideration the difficulty of the studied item.

DIF was simulated to be unbalanced in all of the experiments. That is, the studied items that were simulated to have DIF always indicated DIF against the focal group. Table 2 shows that most studies assessed the magnitude of uniform DIF on power and Type I error (five studies) or just power (one study). To assess power, the difficulty parameters (*b*) of the studied items were increased by some constant for the focal group to indicate small (negligible), moderate, or large uniform DIF. To assess Type I error, the difficulty parameters of the studied items were made equal for the reference and focal groups to indicate no uniform DIF. Only one study investigated the magnitude of non-uniform DIF on power and Type I error. In this case, power involved increasing the discrimination parameter (*a*) of the studied item by some constant for the focal group to represent small, moderate, or large non-uniform DIF. Type I error involved keeping the discrimination parameter of the studied item equal for both groups to represent no non-uniform DIF. Although two studies indicated that "no true DIF was present in the simulated dataset," the authors did not openly state that the difficulty parameters (*b*) for the reference and focal groups were equal.

Table 1: Data Generation and Examinee Conditions

Simulation Study	*Model*	*Data-Type*	*Test Length*	*Sample Size*	*Impact*
Banks & Walker (2006)	3 PL IRT	D	30	$n_F = 250$, $n_R = 1000$ $n_F = 500$, $n_R = 1000$ $n_F = n_R = 1000$	$M_F = M_R = 0$, $SD_F = SD_R = 1$ $M_F = -.5$, $M_R = 0$; $SD_F = SD_R = 1$
Emenogu et al. (2010) Study 2	2 PL IRT	D	25	$n_F = n_R = 2000$	$M_F = M_R = 0$, $SD_F = SD_R = 1$
Falenchuck & Herbert (2009)	2 PL IRT	D	25	$n_F = 500$, $n_R = 3500$ $n_F = 1000$, $n_R = 3000$ $n_F = n_R = 2000$	$M_F = M_R = 0$, $SD_F = SD_R = 1$
Finch (2011a)	3 PL IRT	D	40	$n_F = n_R = 250$ $n_F = n_R = 500$ $n_F = n_R = 1000$	$M_F = M_R = 0$ $M_F = 0$, $M_R = -.5$
Finch (2011b)	3 PL IRT	D	20, 40	$n_F = n_R = 250$ $n_F = n_R = 500$ $n_F = n_R = 1000$	$M_F = M_R = 0$ $M_F = 0$, $M_R = .5$ $M_F = 0$, $M_R = -.5$
Garrett (2009)	Partial Credit	P	20	$n_F = 100$, $n_R = 900$ $n_F = 300$, $n_R = 700$ $n_F = 317$, $n_R = 1183$ $n_F = 355$, $n_R = 845$ $n_F = n_R = 500$	$M_F = M_R = 0$, $SD_F = SD_R = 1$ $M_F = -.5$, $M_R = 0$; $SD_F = SD_R = 1$
Robitzsch & Rupp (2009)	1 PL IRT	D	20, 40	$n_F = n_R = 250$ $n_F = n_R = 1000$ $n_F = n_R = 4000$	$M_F = M_R = 0$, $SD_F = SD_R = 1$ $M_F = 0$, $M_R = .5$; $SD_F = SD_R = 1$ $M_F = 0$, $M_R = -.5$; $SD_F = SD_R = 1$
Rousseau et al. (2006)	2 PL IRT	D	40	$n_F = n_R = 2500$	Not given
Sedivy et al. (2006)	Graded Response	P	9	$n_F = n_R = 500$ $n_F = n_R = 1000$ $n_F = n_R = 4000$	$M_F = M_R = 0$, $SD_F = SD_R = 1$

Note: PL = parameter logistic, IRT = item response theory, D= Dichotomous, P= Polytomous, *n* = sample size, F = focal, R = reference, *M* = mean, *SD* = standard deviation.

Table 2 reveals that the nine studies operationalized MCAR, MAR, or MNAR as they were conceptualized by Rubin (1976); with MCAR simulated most often (six studies), followed by MAR (four studies), and MNAR (six studies). The total exceeds nine because some studies investigated more than one missing data mechanism. The percentage of those with missing data corresponded to the missing data mechanism being investigated. Complete datasets were produced in cases where 0% of the test-takers had missing data. The process allowed researchers to establish a baseline comparison for the incomplete datasets. In order to compare the Type I error rates and power rates for the incomplete and complete datasets, the incomplete datasets had to be made complete using a missing data technique, and both datasets had to be subjected to a DIF analysis. Each of these things is discussed in the next section. Note that Type I error rates are determined by dividing the number of times the studied item was incorrectly flagged for DIF by the number of replications for that condition. Power rates are determined by dividing the number of times the studied item was correctly flagged for DIF by the number of replications for that condition.

Analysis of Conditions

Incomplete datasets were made complete using a given missing data technique. Some studies evaluated more than one technique. According to Table 3 (ignore

Table 2: Data Manipulation Conditions

Simulation Study	*No. Studied Items*	*Magnitude of Uniform DIF*[1]	*Missing Data Technique*	*% Examinees Missing Data*
Banks & Walker (2006)	6 (2 easy, 2 moderate, 2 hard)	None $b_F = b_R$ Moderate $b_F - b_R = .3$ Large $b_F - b_R = .6$	MAR – F omit MNAR – F omit if ability level below difficulty of studied item	0%, 5%, 10%
Emenogu et al. (2010) Study 2	25	None, but no b_F or b_R information given	MNAR – F, R omit contingent on ability group	0%, missing data percentages not given
Falenchuck & Herbert (2009)	25	None, but no b_F or b_R information given	MNAR – F, R omit contingent on ability group	0%, missing data percentages not given
Finch (2011a)	1 (easy, moderate, hard)	None $b_F = b_R$ Moderate $b_F - b_R = .3$ Large $b_F - b_R = .6$	MCAR – F, R omit MAR – F omit MNAR – F, R omit if answered item incorrectly in complete dataset	0%, 5%, 15%
Finch (2011b)	1	None $a_F = a_R$ Small $a_F - a_R = .4$ Moderate $a_F - a_R = .8$ Large $a_F - a_R = 1$	MCAR – F, R omit MAR 1 – F omit MAR 2 – F, R omit if total score at or below 30th percentile MNAR – F, R omit if answered item incorrectly in complete dataset	0%, 10%, 20%, 30%
Garrett (2009)	2	None $b_F = b_R$ Negligible $b_F - b_R = .25$ Moderate $b_F - b_R = .50$ Large $b_F - b_R = .75$	MCAR – F, R omit	0%, 10%, 25%, 40%
Robitzsch & Rupp (2009)	1	None $b_F = b_R$ Negligible $b_F - b_R = .2$ Moderate $b_F - b_R = .4$ Large $b_F - b_R = .6$	MCAR – F, R omit MAR 1 – F omit MAR 2 – F omit if total score at or below 10th percentile, 30th percentile MNAR – F omit if answered item incorrectly in complete dataset	0%, 10%, 30%
Rousseau et al. (2006)	8 items omitted Last 4 items not-reached	Moderate $b_F - b_R = .3$ with 4 omitted items Large $b_F - b_R = 1$ with 4 omitted items	MCAR – F, R, omit	8 items (15%, 25% omit) Last 4 items (15% not-reach)
Sedivy et al. (2006)	1	None $b_F = b_R$ Small $b_F - \mathrm{b}_R = .2$ Moderate $b_F - b_R = .4$ Large $b_F - b_R = .6$	MCAR – F, R omit	0%, 10%, 30%

[1] All of the simulation studies tested for uniform DIF except for Finch (2011b) who tested for non-uniform DIF.

Note: a = discrimination parameter, *b* = difficulty parameter, DIF = differential item functioning, F = focal, R = reference, MAR = missing at random, MCAR = missing completely at random, MNAR = missing not at random.

the bolding for the moment), the most frequently used missing data techniques were zero imputation and listwise deletion (seven studies each), multiple imputation (four studies), and analysiswise deletion (two studies). The remaining missing data techniques were only used once. A variety of DIF detection procedures were applied to the complete and incomplete datasets using the rest score (five studies) or total score/proportion correct of attempted items (two studies) as the matching subtests. Two studies did not give information about how the matching criterions were formed. Some studies evaluated more than one DIF method. In fact, the most commonly used ones were Mantel-Haenszel (six studies), logistic regression (four studies), ordinal logistic regression and SIBTEST (two studies each). The other DIF procedures were only used once.

Recommendations for Practitioners and Researchers

The major results of each study are provided below along with recommendations for the best way to handle missing data when conducting DIF analyses. These suggestions are bolded in Table 3. Banks and Walker (2006) observed inflated Type I error rates for zero imputation under the MAR mechanism, especially when the studied item was easy and 10% of the focal group had missing data. These error rates were greatly reduced under the MNAR mechanism. The researchers were not surprised to find that zero imputation produced larger power rates than listwise deletion given its inflated Type I error rates. The power rates for both missing data techniques increased as the magnitude of uniform DIF against the focal group increased, but decreased as the difficulty of the studied item increased. Listwise deletion was recommended because it produced Type I error and power rates that were similar to the complete datasets.

Emenogu et al. (2010) Study 2 obtained false DIF when zero imputation was used regardless of the matching subtest, while analysiswise deletion did so when total score was the matching criterion. Although listwise deletion did not produce false DIF with either matching subtest, the authors were concerned about the reduction in sample size. Falenchuk and Herbert (2009) obtained false DIF with zero imputation and listwise deletion irrespective of the matching subtest, and also with analysiswise deletion when total score was the matching criterion. Both studies recommended analysiswise deletion with proportion correct of attempted items (number of correct responses out of number attempted) as the matching subtest because no false DIF occurred with this combination.

Finch (2011a) observed that regardless of the DIF procedure, the Type I error rates for zero imputation were inflated when it was used to fill-in missing responses to easy or moderately difficult items under

Table 3: Analysis of Conditions and Recommendations for Practitioners and Researchers

Simulation Study	*Missing Data Technique*	*Matching Subtest*[1]	*DIF Detection Method*
Banks & Walker (2006)	**LD**, ZI	Rest score[2]	SIBTEST
Emenogu et al. (2010) Study 2	**AD**, LD, ZI	Total score **Proportion correct of attempted items**	M-H
Falenchuck & Herbert (2009)	**AD**, LD, ZI	Total score **Proportion correct of attempted items**	M-H
Finch (2011a)	**LD**, **MI**, ZI	Not given	**LR**, **M-H**, **SIBTEST**
Finch (2011b)	**LD**, **MI**, SRI, ZI	Rest score	**CSIB**, **IRTLR**, **LR**
Garrett (2009)	**MI**, Within-Person Mean Substitution	Rest score	**M-H**, OLR
Robitzsch & Rupp (2009)	**LD**, **MICE**, ZI, **Two-Way**, **Two-Way Adjusted**	Rest score	**LR**, **M-H**
Rousseau et al. (2006)	MI, N-R, ZI	Not given	LR, **M-H**, NCDIF
Sedivy et al. (2006)	LD, **LSI**	Rest score	**OLR**, PSIB

[1] Matching subtest always had complete data.

[2] Score on remaining non-studied items.

Note: AD = analysiswise deletion, CSIB = crossing simultaneous item bias test, IRTLR = item response theory likelihood ratio, LD = listwise deletion, LR = logistic regression, LSI = lowest score imputation, M-H = Mantel-Haenszel, MI = multiple imputation, MICE = multivariate imputation by chained equations, NCDIF = non-compensatory differential item functioning, N-R = not-reached, OLR = ordinal logistic regression, PSIB = poly simultaneous item bias test, SIBTEST = simultaneous item bias test, SRI = stochastic regression imputation, ZI = zero imputation.

the MAR mechanism. However, when the data were MCAR or MNAR, these error rates were similar to the complete datasets. Type I error inflation was also observed for zero imputation when 15% of the test-takers missed the studied item. The power rates for zero imputation were similar to the complete datasets under each missing data mechanism. Also, the power rates for all three missing data techniques were higher for larger sample sizes and larger magnitudes of uniform DIF against the focal group, but smaller for harder items. Listwise deletion and multiple imputation were suggested because their Type I error and power rates were comparable to the complete datasets across the DIF methods.

Finch (2011b) obtained Type I error rates for three of the four missing data techniques (excluding listwise deletion) that were larger than the complete datasets across the DIF methods. However, the largest error rates occurred for zero imputation and stochastic regression imputation (similar to regression imputation with random error term) under the MAR1 mechanism, especially as the percentage of focal individuals with missing data increased. These error rates were similar to the complete datasets under the MAR2, MCAR, and MNAR mechanisms. Fairly large Type I error rates were also obtained for zero imputation and stochastic regression imputation when impact was present in large sample size cases. In general, the power rates for all four missing data techniques were somewhat lower than the complete datasets. The power rates increased as the sample size or magnitude of non-uniform DIF against the focal group increased, but decreased as the percentage of individuals with missing data increased. For each DIF method, listwise deletion was the suggested traditional missing data technique, and multiple imputation was the suggested imputation one.

Garrett (2009) found that with each DIF procedure, the Type I error rates for within-person mean substitution (uses mean score for each person) were slightly larger than those for multiple imputation across both studied items. However, in all cases, these error rates were similar to the complete datasets. The power rates for both missing data techniques increased as the magnitude of uniform DIF against the focal group increased, but decreased as the sample size ratios became more disparate or the percentage of persons with missing data increased. Multiple imputation was recommended, possibly due to its slightly better control over Type I errors and slightly larger power rates at higher magnitudes of uniform DIF across the DIF procedures.

Robitszch and Rupp (2009) found inflated Type I error rates (as high as 100%) and limited power rates (as low as 0%) across many conditions in their study; therefore, this information was not reported. The authors did however provide information regarding bias. Positive bias (DIF estimate larger than its parameter on average) occurred using zero imputation under the MAR1 and MAR2 mechanisms, while no bias occurred for the other missing data techniques. Although zero imputation had no bias under the MNAR mechanism, the other missing data techniques had negative bias (DIF estimate smaller than its parameter on average). All of the missing data techniques had no bias under the MCAR mechanism. When bias did occur, it was often higher for larger percentages of examinees with missing data. The authors warned against the use of zero imputation given its tendency to overestimate the amount of DIF present. They also indicated that their study results did not depend on the type of DIF detection method used and considered both to be appropriate.

Rousseau et al. (2006) observed that across missing data techniques, non-compensatory DIF had somewhat inflated false positive rates (falsely identify DIF in items) and fairly large true positive rates (truly identify DIF in items). Logistic regression had somewhat lower false positive and true positive rates than non-compensatory DIF. The false positive rates of both DIF procedures tended to increase as the percentage of examinees with missing data increased, while the true positive rates tended to decrease in this case. The false positive rates of Mantel-Haenszel remained constant at 0% for each missing data technique and percentage of those with missing data. The true positive rates of this DIF procedure remained constant across missing data techniques and percentages of those with missing data, but were significantly lower than the other DIF methods. The authors concluded that the choice of missing data technique should be based on whether it is more important to falsely identify items as DIF or truly do so. Note that non-compensatory DIF assumes that all items except the studied item are DIF-free.

Sedivy et al. (2006) obtained Type I error rates for each missing data technique that were similar to the complete datasets across both DIF methods. The only situations where this was not the case was when Poly-

SIBTEST (similar to SIBTEST but useful for polytomous data) was employed after 30% of the 500 or 1000 examinees had been deleted listwise. In these cases, the DIF effect size parameter could not be estimated because there were not enough observations at each trait level to match the reference and focal groups. The power rates for both DIF procedures tended to be larger when lowest score imputation (similar to zero imputation) was used rather than listwise deletion. These power rates increased as the sample size and magnitude of uniform DIF against the focal group increased, but decreased as the percentage of examinees with missing data increased. Although Poly-SIBTEST had larger power rates than ordinal logistic regression (similar to logistic regression), there were cases where the DIF parameter in Poly-SIBTEST could not be estimated. This occurred when small, moderate, or large DIF was assessed after 30% of the 500 or 1000 examinees had been deleted listwise. Lowest score imputation was the recommended missing data technique and ordinal logistic regression was the suggested DIF method in small sample size cases.

Implications for Practitioners and Researchers

The above literature review has clear implications for practitioners and researchers who are tasked with conducting DIF analyses in the presence of missing data. To begin, practitioners and researchers should avoid zero imputation whenever possible. Zero imputation retains examinees for all DIF analyses by scoring their missing item responses as incorrect. Robitzsch and Rupp (2009) stated that zero imputation is not a "true" imputation method because it is not based on any kind of statistical model. Unfortunately, it is common practice in the field of educational testing and reflects the notion that examinees missing item responses are due to their lack of ability (Ludlow & O'Leary, 1999). Despite the popularity of zero imputation, this review showed that it can lead to inflated Type I errors. For example, Banks and Walker (2006) and Finch (2011a, 2011b) obtained inflated Type I error rates when focal members randomly omitted the studied item (MAR, MAR1) and their missing response was imputed with a zero. These error rates were greatly reduced when focal and reference (MAR2) or focal (MNAR) members randomly omitted the studied item based on their ability level. The Type I error inflation observed with zero imputation was more pronounced for easier items and larger percentages of test-takers with missing data.

Practitioners and researchers should consider using one of the commonly recommended missing data techniques given that they often produce Type I error and power rates similar to the complete datasets as evidenced by Banks and Walker (2006), Finch (2011a, 2011b), and Garrett (2009). Table 3 showed that the most commonly suggested missing data techniques were listwise deletion (four studies), multiple imputation (three studies), and analysiswise deletion (two studies). Individuals who are interested in a quick and easy way to handle missing data should employ listwise or analysiswise deletion. Both of these techniques are user options in SPSS. One of the main drawbacks to listwise deletion however, is reduced sample size which can bias parameter estimates (Graham, 2009). Sedivy et al. (2006) found that in small sample cases, Poly-SIBTEST could not calculate a DIF parameter estimate after listwise deletion because there were not enough observations at each trait level to match the reference and focal groups.

Practitioners and researchers should understand that certain missing data techniques may function better when paired with certain types of matching subtests. Recall that Falenchuck and Herbert (2009) and Emenogu et al. (2010) Study 2 obtained the most valid DIF results when analysiswise deletion was used as the missing data technique and the proportion correct of the attempted items served as the matching criterion. The same could not be said when listwise deletion and zero imputation were used as the missing data techniques, nor when the matching subtest involved the total score. The drawbacks of listwise deletion and zero imputation were previously discussed. With regards to the total score, Emenogu et al. (2010) stated, "… because using the total number of items correct as the matching criterion in MH DIF analyses effectively treats missing responses as wrong, matching on the proportion of items answered correctly out of those attempted may be an appropriate alternative if the assumption is in doubt that the missing data are related to the construct that the test is intended to measure" (p. 460).

Practitioners and researchers should consider that although a range of DIF detection methods have been recommended, only a few of them were suggested more than once. Table 3 showed that Mantel-Haenszel (four studies) and logistic regression (three studies) were the most commonly recommended DIF procedures. Both methods are appropriate for dichotomous items and can detect uniform DIF, while logistic regression can also detect non-uniform DIF. It is also important to note that while some factors can increase the power to detect true DIF, others can decrease it. This review showed that power increases as the magnitude of uniform (non-uniform) DIF against the focal group increases, and also as the sample size increases. By contrast, power decreases as the studied item becomes more difficult and the percentage of examinees with missing data increases (see Banks & Walker, 2006; Finch 2011a, 2011b; Garrett, 2009; Sedivy, et al., 2006). Although these conclusions could have been reached through intuition, the empirical findings were nonetheless insightful.

Suggestions for Future Research

It is evident that additional missing data DIF scholarship is needed. Researchers should consider simulating polytomous data as often as they simulate dichotomous data. This would enable them to take advantage of DIF methods designed for constructed-response data such as Poly-SIBTEST which was used in only one study. Scholars who are interested in simulating dichotomous data should consider applying a 3-PL model given that such a model takes into account the probability that low ability examinees might answer some items correctly because of guessing.

In practically all of the studies, the group of interest was only allowed to omit one item at a time. Future research should consider instances where individuals have missing data on multiple items at once. It would be interesting to determine if DIF results differ depending on the number of items that focal and/or reference members leave blank. Additional research is needed that takes into account the difficulty of the studied item. It is probably safe to assume that as item difficulty increases, the chance of omission also increases, and vice versa.

More research needs to focus on how different missing data techniques react to non-uniform DIF. Finch (2011b) observed less Type I error with zero imputation when assessing non-uniform DIF than when assessing uniform DIF (see Finch, 2011a), especially under the MCAR and MNAR mechanisms. Researchers could benefit from varying the matching subtest instead of relying on the rest score or total score as the matching criterion. Although rest score and total score are commonly used, this review showed that the proportion correct of attempted items could be more appropriate when data are missing.

References

Banks, K., & Walker, C. M. (2006, April). *Performance of SIBTEST when focal group examinees have missing data.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues & Practice, 17(1),* 31-44.

Emenogu, B. (2006). *The effect of missing data treatment on Mantel-Haenszel DIF detection* (Unpublished doctoral dissertation). Ontario Institute for Studies in Education in the University of Toronto.

Emenogu, B. C., Falenchuck, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research, 56(4),* 459-469.

Falenchuk, O., & Emenogu, B. (2006, April). *Differential non-response rates as a source of DIF.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Falenchuk, O., & Herbert, M. (2009, April). *Investigation of differential non-response as a factor affecting the results of Mantel-Haenszel DIF detection.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Finch, H. (2011a). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education, 24,* 281-301.

Finch, H. W. (2011b). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement, 71(4),* 663-683.

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size* (Unpublished doctoral dissertation). Georgia State University.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576.

Herbert, M., Falenchuk, O., & Childs, R. A. (2008, March). *Examination of thin versus thick matching for the Mantel-*

Haenszel DIF procedure in the presence of differential non-response rates in the focal and reference group. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Howard, P. W., & Wainer, H. (1993). *Differential item functioning*. Mahwah, NJ: Erlbaum.

Li, H-H., & Stout, W. (1996). A new procedure for detection of Crossing DIF. *Psychometrika, 61(4),* 647-677.

Ludlow, L. H., & O' Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59(4),* 615-630.

Peng, C. J., & Zhu, J. (2008). Comparison of two approaches for handling missing covariates in logistic regression. *Educational and Psychological Measurement, 68(1),* 58-77.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and Logistic Regression analysis. *Educational and Psychological Measurement*, *69(1),* 18-34.

Rousseau, M., Bertrand, R., & Boiteau, N. (2004, April). *Impact of missing data on robustness of DIF IRT-based and non IRT-based methods.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Rousseau, M., Bertrand, R., & Boiteau, N. (2006, April). *Impact of missing data treatment on the efficiency of DIF methods.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Rupp, A. A., Choi, H. C., & Ferne, T. (2006, April). *Methodological issues in differential item functioning analyses with missing data: A practical illustration with data from the SAIP assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006, April). *Detection of differential item functioning with polytomous items in the presence of missing data.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/dif. *Psychometrika*, *58,* 159-194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27(4),* 361-370.

Citation:

Banks, Kathleen (2015). An Introduction to Missing Data in the Context of Differential Item Functioning. *Practical Assessment, Research & Evaluation*, 20(12). Available online: http://pareonline.net/getvn.asp?v=20&n=12

Author:

Kathleen Banks
348 College of Education
Middle Tennessee State University
Murfreesboro, TN 37130

kathleen.banks [at] mtsu.edu