Running Head: Application of Propensity Model in DIF Studies

Application of Propensity Models In DIF Studies To Compensate For Unequal Ability Distributions

Kevin Joldersma and Dan Bowen

Measurement Incorporated, Durham, NC

Keywords: DIF, differential item functioning, propensity models, language testing, linguistic factors of DIF

Abstract

In this paper, we seek to advance the application of propensity based models in psychometric research on differential item functioning. These models are commonly used in econometrics and medical research to account for differing distributions between reference and focal groups. The research study is conducted upon a large-scale state-wide assessment of language arts literacy that was translated from English into Spanish. To provide the necessary statistical evidence of equivalence between the two versions of the test better methodology for quantitatively identifying possible biased items needs to be established. It is hypothesized that by using propensity score matching could be used to control for the disparate distributional differences that exist between the populations of examinees who received each version of the test.

Introduction

In the development of any psychological, educational, or licensure test, an essential consideration is ensuring that the test is fair to all test takers and that bias against any segment of the population has been minimized. The results from tests and assessments are often used as a guide for understanding the test taker's ability or rank, or for decisions related to advancement, placement, and licensure. Inferences are made based on test results because it is assumed that tests yield valid information about test taker ability on a given construct. These assumptions can have a range of ramifications: students pass or fail, treatments are recommended or not, drivers are granted or denied licenses. Test scores have meaningful implications for individuals and society; therefore, it is important that possible threats to the basic assumptions of a test be investigated.

One potential threat to test validity is test bias. According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), test bias "is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups." (p. 74). There are several potential causes for measurement bias; one frequent underlying cause is that the test is differentiating between test-takers based on a characteristic that is irrelevant to the construct being measured. No test can ever be completely free from bias. Crocker and Algina (2006) posit that test scores will always be subject to sources of construct irrelevant variance. However, when the distributions of the irrelevant construct differ substantially between two subgroups of examinees, one subgroup of test takers may be unfairly advantaged. When this occurs, it is referred to as "differential item functioning" (DIF). When present, DIF contributes reduced validity of test scores. By

performing statistical analyses (DIF analyses) on the items of a test, problematic items that may be measuring differently for different groups of test takers and test bias can be identified and addressed (Camilli & Shepard 1994).

Increasingly, to accommodate their diverse populations, test developers and users of tests are translating their exams into multiple languages. This has opened the door for a range of validity issues because to simply translate a test from one language to another does not provide the necessary evidence to support the validation of its use. Hambleton (2000) lists 22 guidelines for adapting and translating tests into different languages. As part of the responsibilities of test developers and publishers, he states that "statistical evidence of equivalence of questions for all intended populations" (p. 168) should be provided. To provide evidence of equivalence, studies to identify possible item bias are recommended. There are a myriad of differential item functioning (DIF) statistical methods available to researchers to address bias; however, as described below, most are inadequate for the unique efforts needed to provide evidence to support the usage of test translations and adaptations.

Current DIF Methodologies

It is important to note the difference between DIF and item bias. Zumbo (1999) defines the occurrence of item bias as "when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose." However, DIF is defined as "when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the

underlying ability that the item is intended to measure." Thus, not all items that show DIF are biased, but all items that are biased show DIF.

Over the years many different procedures have been used to identify DIF. Some of the early methods were based upon classical test theory (CTT) and analysis of variance (ANOVA) (Camilli and Shepard, 1994). These methods rely on p-value differences between the two groups, without adjusting for latent ability on the construct, and thus are seriously flawed. After all, items that are the most discriminating between two groups will show the largest p-value differences. Furthermore, an item may seem to favor the reference group when looking at p-value differences between the two groups, but when the groups are broken down into ability levels, the item may actually favor the focal group, a testing version of Simpson's Paradox. Thus, CTT and ANOVA methods are no longer recommended for use in identifying DIF. More modern methods are based off of item response theory (IRT), logistic regression, and the Mantel-Haenszel chi-square test.

IRT involves using the number of correct responses to estimate an examinee's ability via a one-, two-, or three-parameter logistic model. The estimates can then be compared across focal and reference groups (Camilli & Shepard, 1994). Once scores are standardized they can be displayed in an Item Characteristic Curve (ICC), where the X-axis represents ability level and the Y-axis represents the probability of getting item correct. DIF exists if the ICCs of each group are different; in other words, the ICCs of the focal and reference groups represent unique populations (Crocker & Algina, 2006). An advantage of using IRT is it is capable of identifying both uniform and non-uniform DIF. Uniform DIF occurs when the ICCs of the reference and focal groups do not intersect. Conversely, non-uniform DIF occurs when their ICCs do intersect.

Swaminathan and Rogers (1990) recommended using logistic regression as a method for identifying DIF. The benefits of using logistic regression include the ability to identify both uniform and non-uniform DIF. Conceptually, it is a procedure where group, ability, and a group/ability interaction are used to calculate the probability of a correct or incorrect answer to an item. The group variable represents group membership and ability typically represents total score on a given construct. If the group variable is statistically significant, this indicates that the probability of getting the item correct is different for each group, after controlling for ability. Ability is almost always statistically significant because unless it is a misfit item, the examinees with higher ability will have a greater probability of answering an item correctly than examinees with lesser ability. Finally if the group/ability interaction is significant, non-uniform DIF is present.

Mantel and Haenszel (1959) introduced their extension of the Chi-square test to address one of the limitations of retrospective studies of rare diseases. They argued that many times forward studies of rare diseases were simply unfeasible due to the number of people needed to undertake them and the cost involved. Retrospective studies, on the other hand, could be much cheaper and accomplished with fewer people in groups. Mantel and Haenszel's goal was "to reach the same conclusions in a retrospective study as would have been obtained from a forward study" (p. 722).

Holland and Thayer (1988) adapted the Mantel-Haenszel procedure for identifying DIF. The Mantel-Haenszel procedure is essentially a 2x2xK contingency table (Camilli & Shepard, 1994; Mantel & Haenszel, 1959). Data are entered into there contingency table where there are two categorical variables with K ordered levels. The null hypothesis is that there is no linear trend among the ordered categories. The contents

of the columns of the table represent whether the item was answered correctly (1) or incorrectly (0), and the row contents represent whether the examinees are members of the focal or reference groups. The focal group, typically considered to be the minority group, is the group being investigated to determine whether their responses to an item are an accurate representation of their standings on a construct. The reference group, typically the majority group, is the group which is being used as a standard for comparison to the focal group. Examples of focal and references groups could be male and female, economically disadvantaged and non-economically disadvantaged, or Hispanic and Caucasian, respectively. The Mantel-Haenszel procedure tests the assumption of a uniform effect (i.e the measure of an effect is consistent or homogeneous across levels) and that the total score provides an overall measure (through 'comparing the comparable'). If the items are stratified and compared to the overall or whole score (ratios) it can be determined whether or not there is a difference in effect among strata. The Mantel-Haenszel procedure therefore provides a measure similar to a pooled odds-ratio or rate difference. The null hypothesis is that the pooled odds ratio, or relative risk, equals one (there is no association between rows and columns).

Problem Statement

There are unique problems for educational researchers in attempting to validate the inferences made from the administration of test translations. In most cases the population that has received the accommodation to adapt the test to their native language will be much smaller than the population taking the test in its original language. Most modern methods for identifying DIF produce spurious results if the samples sizes are

small. In their research on Mantel-Haenszel, Mazor, Clauser, and Hambleton (1992) state its two major benefits compared to IRT and Logistic Regression: 1) it requires less computing power and 2) it can be used with small sample sizes. A normal test translation/adaptation is not likely to have more than a few hundred examinees; IRT and Logistic Regression require many more examinees to produce meaningful results. Thus, of the three methods, only Mantel-Haenszel could feasibly be used, in most test adaptations.

There is a caveat to this finding. Mantel-Haenszel's statistical power decreases as the distributions of the sample sizes of the reference and focal groups become more and more unequal (Mazor et al., 1992). In their Monte Carlo study, Herrera and Gomez (2008) tested the influence of unequal sample sizes on the detection of DIF using the Mantel-Haenszel procedure. As the differences between the number of cases in the reference and focal groups got larger, the detection rate of DIF went down. In high stakes state assessments where an exam has been adapted to another language the sample size of the reference group will likely have over 100 times the amount of examinees as the focal group.

Another disconcerting dilemma is the ability distributions of the reference and focal groups. Mantel-Haenszel's power decreases as the ability distributions become more and more disparate (Herrera and Gomez, 2008; Mazor et al., 1992; Muniz, Hambleton, Xing, 2001; Narayanan and Swaminathan, 1994). In high stakes state assessments test adaptations will likely be accommodations for a population of examinees with an ability distribution that is vastly different than the normal population of test takers. If Mantel-Haenszel is to be used to compare the reference and focal groups

in assessing DIF in test adaptations, methodology will have to be identified and employed to control for the disparate distributional differences of the two populations.

Empirical Investigation

Research Question

Can propensity score matching be used to control for the disparate distributional differences that exist between the populations of examinees who received each version of a test translation in an effort to more accurately identify DIF?

Propensity Score Matching as a Solution

Propensity score matching was originally used in the Biometric and Econometric fields. Rosenbaum and Rubin (1983) applied it to observational studies in an effort to remove the bias between the characteristics of the treatment and control groups. For instance, if a researcher is studying the effects of a method of enhancing reading skills in children, then he or she would want to know that the treatment is the cause of group differences, not other confounding factors such as latent reading ability, ethnicity, gender, etc. Large differences between the two groups would have to be reduced to ensure there is no treatment selection bias. For each student a probability of receiving the treatment or control would be calculated - with a logistic regression equation - based on their various characteristics and any variables the researcher deemed appropriate. The propensity score would represent the probability of being in the treatment group. Individual students in the treatment and control group are then matched on their propensity scores and the two groups may be compared with all the group differences being controlled for.

There are different algorithms that may be used to accomplish matching, including greedy propensity score matching, optimal propensity score matching, and Heckman et al's kernel-based matching (Guo and Fraser, 2010). Our research used greedy matching, because it allowed for us to use all of our data from the small minority population. Furthermore, since we had a large range of overlapping propensity scores and a sufficient number of cases in the minority group, major assumptions of greedy propensity score matching were satisfied (Rosenbaum, 2002; Parsons, 2001). Thus greedy matching will be the focus of our discussion. A more detailed discussion of the other matching algorithms can be found in "Propensity Score Analysis" by Guo and Fraser.

The greedy algorithm is used to match the propensity scores of one case from the treatment group to one case in the control group, within a certain pre-specified tolerance. The match that is made for any given case is always the best available match and once that match has been made it cannot be reversed. The rest of the matches follow the exact same pattern. At every step in the process amongst the remaining cases from the treatment and control groups, best available matches are made and those matches cannot be reversed (Parsons, 2001). After the matching has been completed, comparisons of the treatment and control groups may be made with the specified confounding variables controlled.

Methodology

The study is a two-part study. The first step was to run traditional DIF analyses using the Mantel-Haenszel procedure without any compensation for distributional differences. The next step was to use a propensity score matching greedy matching

algorithm (Parsons 2001) to account for the distributional differences. Three different models were used for matching. Table 4 illustrates the efficacy of the matching techniques. Next, the Mantel-Haenszel procedure was rerun with the reference groups being the matched samples and the focal group being the population of examinees who received the Spanish version of the test.

Data

The data used in this study are summarized in Table 1 and were acquired from a large-scale state-wide assessment of Language Arts Literacy that was translated from English into Spanish. Initial distributional differences between the two populations are summarized in Table 2. The test administration was for grades 5-8. Initial DIF analysis was carried out by analyzing the responses of the 34,080 to 38,539 examinees who took the English version of the exam and the 554 to 713 examinees who took the Spanish version of the exam.

Analysis

Analysis was performed using the Mantel-Haneszel procedure for identifying DIF and the initial results are summarized in Table 3. Using the ETS classification system for identifying DIF, we found that approximately one quarter of the items (37 out of 144) were classified as C, meaning they show large DIF and should be reviewed carefully for bias or removed from the test. Since the initial analysis did not control for the disparate distributional differences, the Mantel-Haenszel results are possibly spurious because of the massive disparities in ability levels and population sizes of the reference and focal groups.

Table 1: Data Summary

Grade	Language	Examinees	Average Score	St. Dev.
5	English	35,472	41.4	10.7
	Spanish	554	27.4	9.8
6	English	34,080	41.4	10.6
	Spanish	660	26.3	9.6
7	English	35,093	43.3	10.7
	Spanish	713	29.6	9.0
8	English	38,539	49.7	9.6
	Spanish	663	37.0	9.9

Table 2: Initial Distributional Differences

	Model 0	
Grade	Mean Diff.	St. Dev. Diff.
5	14.1	0.92
6	15.1	1.05
7	13.8	1.75
8	12.6	-0.24

Table 3: Initial MH results: ETS classifications

	Classification	Grade 5	Grade 6	Grade 7	Grade 8	Total
Model 0	A	21	24	19	19	83
	B	8	2	8	6	24
	C	7	10	9	11	37

Models

Model 0 was used for the control model. It was the initial Mantel-Haenszel analysis without controlling for the distributional differences. Model 1 was the omnibus model. Greedy propensity score matching was conditioned upon the item response strings, meaning students in the focal group were matched up to students in the reference group whose item response string most closely resembled their own. Model 2 only used one variable—total score—to match using propensity scores, meaning students in the focal group were matched with one other student in the reference group who received the

exact same total score they did. Model 3 matched with total score, and added two demographic variables: gender and economic status.

Results

Once the propensity score matching was completed, the initial distributional differences were controlled. Table 4 shows the new characteristics of the data sets of each of the three models in comparison to the original data and Table 5 shows the distributional differences for each of the four models at each grade level.

The initial distributional differences (model 0) showed vastly different populations; however, after propensity score matching, the number of examinees in each data set is exactly the same, and the differences in ability levels of the two data sets are almost non-existent. The examinees matched in model 1 have total score averages and standard deviations that are much more consistent with populations of the same ability level than the original model. Unsurprisingly, the examinees matched in models 2 and 3 have total score averages and standard deviations that are almost exactly the same.

Table 4: Sample Characteristics by Model

Model	Grade	Language	Examinees	Average Score	St. Dev.
0	5	English	35,472	41.4	10.7
		Spanish	554	27.4	9.8
	6	English	34,080	41.4	10.6
		Spanish	660	26.3	9.6
	7	English	35,093	43.3	10.7
		Spanish	713	29.6	9.0
	8	English	38,539	49.7	9.6
		Spanish	663	37.0	9.9

Table 4: Sample Characteristics by Model (Continued)

Model	Grade	Language	Examinees	Average Score	St. Dev.
1	5	English	531	26.9	10.2
		Spanish	531	27.5	9.9
	6	English	606	26.6	9.9
		Spanish	606	26.5	9.7
	7	English	671	28.9	9.4
		Spanish	671	29.7	9.1
	8	English	610	35.5	10.8
		Spanish	610	37.5	9.9
2	5	English	552	27.4	9.8
		Spanish	552	27.4	9.8
	6	English	655	26.4	9.6
		Spanish	655	26.4	9.6
	7	English	696	29.6	9.0
		Spanish	696	29.6	9.0
	8	English	658	37.1	9.8
		Spanish	658	37.1	9.8
3	5	English	552	27.4	9.7
		Spanish	552	27.4	9.8
	6	English	655	26.4	9.6
		Spanish	655	26.4	9.6
	7	English	696	29.6	9.0
		Spanish	696	29.6	9.0
	8	English	658	37.1	9.8
		Spanish	658	37.1	9.8

Table 5: Distributional Differences in Total Score by Model

	Model 0		Model 1		Model 2		Model 3	
Grade	Mean Diff.	St. Dev Diff.	Mean Diff.	St. Dev Diff.	Mean Diff.	St. Dev Diff.	Mean Diff.	St. Dev Diff.
5	14.0	0.9	-0.6	0.3	0.0	0.0	0.0	-0.1
6	15.1	1.0	0.1	0.2	0.0	0.0	0.0	0.0
7	13.7	1.7	-0.8	0.3	0.0	0.0	0.0	0.0
8	12.7	-0.3	-2.0	0.9	0.0	0.0	0.0	0.0

Once the initial distributional differences are ameliorated by means of propensity score matching, we began by analyzing the p-values and their correlations between the two groups. The first, simple analysis was done as a first pass at evaluating whether DIF

still existed between the two groups. Table 6 displays the correlations of the p-values of the students in the matched samples. This is another proxy to indicate ability matching of reference and focal groups. The correlations of p-values before any adjustment ranged from ~.07 to ~.44. After PSM, the correlations range from near perfect (~0.98) to relatively weak (~.32). Model 1 has near perfect p-value matching in all grades. This makes sense due to all of the items being modeled by the propensity scoring algorithm. Models 2 and 3 do show, across all grades, an increase in the correlation of p-value between target and reference groups. We speculate that this is due to the examinees being matched and the distributions of these models becoming more similar in shape.

Table 6: P-value correlation by Model

Grade	Model	Correlation
5	O	0.4478
5	1	0.9896
5	2	0.5956
5	3	0.6282
6	O	0.2588
6	1	0.9844
6	2	0.3215
6	3	0.3830
7	O	0.0734
7	1	0.9830
7	2	0.3972
7	3	0.4045
8	O	0.4261
8	1	0.9936
8	2	0.5242
8	3	0.5075

After the new data sets had been created using the propensity score matching, analysis was performed using the Mantel-Haneszel procedure for identifying DIF and the results are summarized in Table 7. Model 1, conditioned upon item response

string, essentially functioned to produce results where DIF would be impossible to identify. Out of 144 items 0 were classified as showing moderate DIF. This is likely due to there being such a large pool of examinees in the original English group that Spanish examinees could be matched with an English examinee with a very similar item response string, thus making DIF nearly impossible to identify. Similarly to the traditional Mantel-Haenszel procedure, model 2 was conditioned upon total score. Out of 144 items model 2 assigned the exact same DIF classification as the original model 77% of the time. Finally, model 3 - conditioned upon total score, gender, and economic status - assigned the exact same DIF classification 82% and 90% of the time with the original model and model 2, respectively. It is unknown whether the differences between models 0, 2, and 3 are statistically significant, or due to chance.

Table 7: ETS Classifications by Model and Grade Level

	Classification	Grade 5	Grade 6	Grade 7	Grade 8	Total
Model 0	A	21	24	19	19	83
	B	8	2	8	6	24
	C	7	10	9	11	37
Model 1	A	36	36	36	35	144
	B	0	0	0	0	0
	C	0	0	0	0	0
Model 2	A	21	18	20	21	80
	B	11	8	5	4	28
	C	4	10	11	11	36
Model 3	A	21	24	20	19	84
	B	10	2	6	7	25
	C	5	10	10	10	35

Table 8 shows the agreement rates of the models in identifying Class C items. To save space, the results for model 1 have not been included; after all, model 1 did not identify one Class C item. Model 0 agreed in identifying Class C items with model 2 66%

of the time and with model 3 64% of the time. Not surprisingly given their similarities, model 2 and model 3 agree in identifying Class C items 92% of the time.

Table 8: ETS Class C items identification agreement rates

	Grade	Model 0	Model 2	Model 3
Model 0	5	-	57%	50%
	6	-	82%	82%
	7	-	43%	46%
	8	-	83%	75%
	Total	-	66%	64%
Model 2	5	57%	-	80%
	6	82%	-	100%
	7	43%	-	91%
	8	83%	-	91%
	Total	66%	-	92%
Model 3	5	50%	80%	-
	6	82%	100%	-
	7	46%	91%	-
	8	75%	91%	-
	Total	64%	92%	-

Discussion

When we initially began using PSM as an effort to account for the effects of disparate populations, we were advised to try numerous variables to control for construct irrelevant differences. Thus, in our model building approach, we began with an omnibus model. Then we took a step back to build the model where we used only total score as a conditioning variable in an effort to parallel traditional Mantel-Haneszel procedures. Finally, we used statistically viable demographic variables to see if any change in information yielded would be produced.

As stated previously, the omnibus model was conditioned on the response string of the examinees. Due to the large number of examinees available from the reference

group the response strings of the match examinees were nearly the same. Consequently, only an item exhibiting extraordinarily large differences in behavior between reference and focal would be flagged by Mantel-Haneszel procedures following PSM under the omnibus model. We strongly recommend against the approach because this type of model can actual work to obscure DIF that would (and should) normally be detected.

The second model for PSM is the closest analog to a traditional Mantel-Haneszel approach. The only difference in this case is that the focal group and their performance are not overwhelmed by the number of examinees in reference group. Despite having controlled for the large distributional difference in population shape and size, the findings are quite consistent with a traditional, non-matched Mantel-Haneszel analysis. Although 66% of the Class C items were flagged in common between the original Mantel-Haneszel analysis and model 2, one cannot help but wonder why the other 44% did not overlap. Model 2 identifies one fewer items as a Class C item than the original Mantel-Haneszel analyses. However, different items are flagged as having some level of DIF.

The third model agrees in Class C item 64% of the time with the original Mantel-Haneszel analyses. Similar to model 2, model 3 flags about the same number of items as Class C, but different items than the original Mantel-Haneszel analyses. The addition of the demographic variables does not yield any large difference in which items are flagged for DIF between models 2 and 3. Modeling different demographic variables may prove otherwise.

We believe that PSM is a tool which allows us to do two beneficial things: condition the DIF analyses upon multiple variables and control for disparate distributional characteristics. These are two extremely important advantages when

working with populations with known differences. Having said this, our results are only a first step. We are left with a quandary—how to tell whether a model has improved our understanding of or identification of DIF between the reference and focal populations. We have answered the first half of our initial question in that PSM clearly can control for large distributional differences and examinee population sizes. In future research, we plan to do additional testing of propensity score matching for DIF with simulation models. Thus, we can test DIF under different conditions and determine what distributional differences and the degree to which the can be accounted for with PSM to refine our identification of problematic items in smaller groups.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. (Vol. 4) Thousand Oaks, CA: Sage Publications.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Mason, OH: Thompson-Wadsworth.

Guo, S., & Fraser, M. F. (2010). *Propensity score analysis: statistical methods and applications.* Thousand Oaks, CA: Sage.

Herrera, A. N., & Gomez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity: International Journal of Methodology, 42*, 739-755.

Hambleton, R. K. (2000). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17, 164-172.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspect of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effects of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, 52, 443-451.

Muniz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115-135.

Parsons, L. S. (2001). Reducing bias in a propensity score matched-pair sample using greedy matching techniques. [SAS SUGI paper 214-26]. Proceedings of the 26th annual SAS Users' Group International Conference, Cary, NC: SAS Institute.

Rosenbaum, P. R. (2002). Observational studies (2nd ed.). New York: Springer.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41-55.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.

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

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-types (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.