9. ITEM ANALYSIS

CHAPTER NINE

Introduction to Psychometric Evaluation

Item analysis is often conducted to evaluate the psychometric properties of individual items on a test form. In accordance with classical test theory (CTT), items are evaluated in terms of item difficulty and item discrimination.[1] Option or distractor analysis is often conducted to further investigate the discrimination power of the item. Further, differential item functioning is examined to flag items with potential bias.

ITEM DIFFICULTY

In CTT, item difficulty, also known as item p-value, is often quantified in terms of the mean item score or the percentage of students who answer the item correctly. It can be computed using the following equation

(9.1)

$$p-value_j = \frac{\sum_{i=1}^{I} X_{ij}}{N} =$$

where X_{ij} is the item score for student j on item I, and N is the total number of students answering the item. Theoretically, p-values range from 0 to 1. A higher p-value means that more students have answered the item correctly, and thus the item is judged to be easier. Vice versa, a lower p-value means that fewer students have answered the item correctly, and thus the item is judged to be more difficult. Usually the p-value falls within the range of 0.25 to 0.9. An item which falls outside this range could be very easy (p-value>0.9) or very difficult (p-value<0.25). The p-value is group dependent. It may vary as the group ability changes. The p-value may be higher for a high ability group whereas it may get lower for a low ability group. This index is not comparable across different test administrations.

ITEM DISCRIMINATION

Item discrimination is another measure that is often used to quantify the item property with the CTT framework. In general, items that can distinguish high ability from low ability students are considered good

1 Novick, M.R. (1966) The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3,1-18.

items. Item discrimination can be quantified in terms of the item-total correlation. This index describes the relationship between students' performance on an item and their performance on the overall test. The point-biserial correlation is a specific type of item-total correlation for dichotomous items. The point-biserial correlation can be computed as follows.

$$p-bs_j = Corr(X_{ij}, X_{i.}) = \frac{Cov(X_{ij}, X_{i.})}{SD(X_{ij})SD(X_{i.})} \quad (9.2)$$

The numerator in the equation is the covariance between the total test scores and the item scores, while the denominator is the product of the standard deviation of the item scores and the test scores respectively. High point-biserial correlations indicates that students who answer an item correctly also have high total test scores and vice versa. There is no consensus related to the cut value for the point-biserial correlation to evaluate item quality. Usually items with point-biserial correlations larger than 0.25 are considered acceptable. Items with a point-biserial correlation smaller than 0.1 are often deemed as not discriminating enough to distinguish between the high and low achievers. Items with a point-biserial correlation between 0.1 and 0.25 are often flagged for further investigation.

OPTION/DISTRACTOR ANALYSIS

Option/distractor analysis shows the percentage of students choosing each option in a dichotomous item. In addition, often the percentage of missing answers is also included in the analysis to get a general idea of the percentage of students who may skip the item. If the missing percentage is larger than 5%, caution should be exercised to scrutinize the item more carefully. This percentage is computed for all students and the subgroups of students. Usually three subgroups are used in such an analysis by dividing students into three categories of ability levels. The total test scores can be used to categorize students into low, middle, and high ability groups. For an item with good discriminating power, it is expected that more high ability students will choose the correct option while the low ability students would be attracted by the distractors which often represent different types of misconceptions.

The following two tables (Tables 9.1 and 9.2) contrast two items. The option highlighted in red indicates the key of the specific item (the correct answer for that question). The numbers in the option point-biserial are the highest values. It is expected that the key option should have the highest option point-biserial correlation. The point-biserial correlation with the item is 0.58 for Item 1 in Table 9.1, while that for Item 2 in Table 9.2 is 0.05. For Item 1, it is evident that the majority of the high and middle ability students chose Option D, which is the key of this item, while the majority of the low ability students chose Options B and C. The p-value of this item is 0.64.

Table 9.1 Option Analysis for an Item with Good Discrimination.

		Option-A	Option-B	Option-C	Option-D	Missing
Item 1	Total	11%	18%	7%	64%	0%
	Low Ability	20%	40%	35%	5%	0%
	Middle Ability	16%	27%	11%	46%	0%
	High Ability	4%	3%	1%	92%	0%
	Option Point-Biserial	-0.17	-0.33	-0.27	0.52	

Table 9.2 Option Analysis for an Item with Low Discrimination.

		Option-A	Option-B	Option-C	Option-D	Missing
Item 2	Total	21%	22%	51%	5%	1%
	Low Ability	31%	19%	33%	15%	1%
	Middle Ability	23%	23%	49%	5%	0%
	High Ability	13%	22%	62%	2%	1%
	Option Point-Biserial	-0.16	-0.02	0.20	-0.13	

On the other hand, the point-biserial correlation for Item 2 is low, thus this item has little discrimination. Regardless of the ability levels, all students are attracted by Option C while Option B is the key. This is reflected in that Option B is the key but has a negative option point-biserial correlation, but Option C has the highest positive option point-biserial correlation. This item is a difficult item with a p-value of 0.22. Such occurrences are rare but not unexpected.

Further, the option point-biserial correlation may present more information about the quality of an item. When computing the option point-biserial correlation, each option would be treated as the correct answer while all other options would be scored as incorrect regardless of the key. It is expected that the correct option will have positive option point-biserial correlation while other options or distractors will have negative option point-biserial correlations. In Table 9.1, the option point-biserial correlations for distractor options A, B, and C are negative while that for the key is positive, as expected. For Item 2 in Table 9.2, the option point-biserial correlation for the correct option is close to 0 while the option point-biserial for distractor Option C is positive, which raises a flag for further investigation.

DIFFERENTIAL ITEM FUNCTIONING

To further investigate the item psychometric properties, differential item functioning (DIF) is conducted. DIF refers to the unexpected differences in students' performance on an item between two groups after they are matched on their ability measured in the test. This conditional comparison is often conducted between two groups; one is called the reference group while the other is called the focal group. The reference group refers to individuals for which the test was expected to favor, most often related to the white or male in educational assessment while a focal group refers to individuals who are at risk or being disadvantaged by the test, most often related to the non-white or female. The following graph illustrates the presence of DIF in an item.

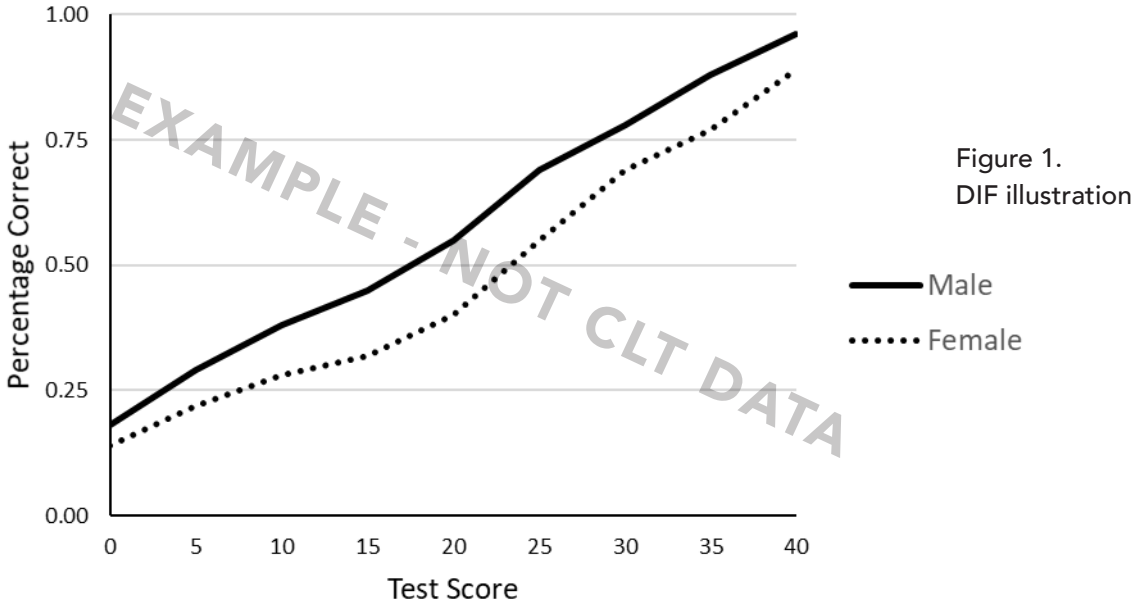

Figure 1.
DIF illustration

Mantel-Haenszel statistics and ETS categories are two commonly used DIF indexes. Mantel-Haenszel alpha can be computed using equation

$$\alpha_{MH} = \frac{\sum_{k=1}^{m} \frac{B_k C_k}{T_k} \frac{A_k D_k}{B_k C_k}}{\sum_{k=1}^{m} \frac{B_k C_k}{T_k}} = \frac{\sum_{k=1}^{m} \frac{A_k D_k}{T_k}}{\sum_{k=1}^{m} \frac{B_k C_k}{T_k}} \quad (9.3)$$

where cell frequencies represented by A, B, C and D in the above equation can be related to Table 9.3.

Table 9.3 Cross-Tabulation Table for Item Responses and Group Membership.

		Item Response Y		
		1	0	Total
Group	Reference	Ak	Bk	nRk
	Focal	Ck	Dk	nFk
	Total	n1k	n0k	Tk

Further, $\hat{\lambda}_{MH} = \ln(\hat{\alpha}_{MH})$ and $D-DIF = -2.35\hat{\lambda}_{MH}$. Items are classified into one of the three DIF categories

Category A: Negligible DIF, no contrast group favored;

Category B: Moderate DIF, one contrast group is slightly favored by the studied item;

Category C: Large DIF, one contrast group is strongly favored by the studied item.

The presence of DIF indicates only that the students with equal ability from different subgroups have an unequal probability of responding to an item correctly. Items in category B and C are flagged and should be carefully examined for potential bias against a particular group.

DIF does not necessarily mean that an item is biased. An item is biased if it measures attributes irrelevant to the intended construct or is somehow a less acceptable measure of the construct for one subgroup. The results of DIF analyses provide a convenient starting point for the study of item bias. Statistical bias does not imply the item is unfair. Expert review of item content is needed.

Summary of Item Analysis Results for CLT Tests

All analyses introduced in the above section were conducted for the April 2018 CLT administrations including test forms 1517 and 1618.

ITEM DIFFICULTY

Item p-values were computed for each item in both forms. Table 9.4 presents the descriptive statistics of p-values for all items in these two forms. The mean p-values ranged from 0.44 to 0.63, with the Quantitative Reasoning test in Form 1618 having the lowest mean p-values. The number of items with p-values smaller than 0.25 are 2 for each section of Form 1517, while those for Form 1618 are 0, 2, and 6 for the Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning respectively. On the other hand, there are 4 items in the Grammar/Writing section of Form 1517 and 1 item in the Grammar/Writing section of Form 1618 that have p-values larger than 0.9. All other sections in Form 1517 and 1618 do not have items with p-values larger than 0.9. The p-value for individual items can be found in Appendix A1.

Table 9.4 Descriptive statistics of P-values.

Form	Subject	N	Mean	SD	Minimum	Maximum	P<0.25	p>0.9
1517	Verbal Reasoning	40	0.56	0.16	0.22	0.86	2	0
	Grammar/Writing	40	0.62	0.24	0.13	0.95	2	4
	Quantitative Reasoning	40	0.52	0.17	0.23	0.85	2	0
1618	Verbal Reasoning	40	0.63	0.16	0.30	0.87	0	0
	Grammar/Writing	40	0.58	0.21	0.16	0.94	2	1
	Quantitative Reasoning	40	0.44	0.18	0.15	0.86	6	0

Note: The last two columns represent the number of items in each section in each form that fall in the range of the indicated values.

ITEM DISCRIMINATION

The point-biserial correlations are summarized in Table 9.5. The mean point-biserial correlations ranged from 0.34 to 0.44 with the largest value of 0.62 in the Grammar/Writing section in Form 1618 and the smallest value of 0.01 in the Quantitative Reasoning section in Form 1618. As items with values ranging from 0.1 to 0.25 raise a warning flag, and those with values less than 0.1 also raise a red flag, such information is summarized as well. Only 4 items had a point-biserial correlation smaller than 0.1 with the lowest value of 0.01 in the Quantitative Reasoning section of Form 1618. The point-biserial for individual items can be found in Appendix A2.

Table 9.5 Descriptive statistics of point-biserial correlation.

Form	Subject	N	Mean	SD	Minimum	Maximum	pbs<0.1	0.1<pbs<0.25
1517	Verbal Reasoning	40	0.34	0.10	0.05	0.48	2	5
	Grammar/Writing	40	0.34	0.10	0.09	0.58	1	5
	Quantitative Reasoning	40	0.39	0.09	0.11	0.53	0	2
1618	Verbal Reasoning	40	0.44	0.11	0.20	0.59	0	4
	Grammar/Writing	40	0.41	0.11	0.19	0.62	0	4
	Quantitative Reasoning	40	0.37	0.14	0.01	0.57	1	7

Note: The last two columns represent the number of items in each section in each form that fall in the range of the indicated values.

OPTION/DISTRACTOR ANALYSIS

As illustrated above, option/distractor analysis further demonstrates item performance in different ability groups. For test security reasons, this information cannot be summarized in this document. In general, the information collected in these analyses further cross-validated what has been observed and summarized above and provided more detailed information about which option may be the potential cause for the low discrimination in items that have been flagged. Option/distractor analysis and option point biserial correlations for each individual item can be provided to relevant stakeholders of CLT upon request with the signing of a confidentiality agreement.

DIFFERENTIAL ITEM FUNCTIONING

Two types of DIF analyses were conducted, one for gender and the other for race. For gender DIF analyses, the male group was designated as the reference group while the female group as the focal group. For the race

DIF analysis, the White student group was treated as the reference group while the Non-White student group was treated as the focal group. Students with missing group indicators were excluded from DIF analyses. difR package (Magis, Beland, Tuerlinckx, & De Boeck, 2010) was used with all the default settings.[2] The number of items flagged with DIF for each section of the CLT is summarized in Table 9.6. In general, the majority of the items in each section were classified with Category A DIF, which is negligible. Usually, items with Category C DIF require further scrutiny of item content. More items were flagged Category C DIF in Form 1618. Some items favored the reference groups while other items favored the focal groups. The detailed information about DIF analysis results for individual items can be found in Appendix A3 to A8.

Table 9.6 Differential Item Functioning

Form	Subject	Grouping variable	N	Number of Items in Each ETS Category		
				A	B	C
1517	Verbal Reasoning	Gender	40	34	3	3
		Race	40	35	3	2
	Grammar/ Writing	Gender	40	33	2	5
		Race	40	29	8	3
	Quantitative Reasoning	Gender	40	35	5	0
		Race	40	28	7	5
1618	Verbal Reasoning	Gender	40	24	6	10
		Race	40	25	10	5
	Grammar/ Writing	Gender	40	29	9	2
		Race	40	21	10	9
	Quantitative Reasoning	Gender	40	25	8	7
		Race	40	25	8	7

2 Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous different item functioning. *Behavior Research Methods,* 42, 847-862.