# Validating a new instrument for measuring computer competence in Lebanese University (Pedagogy Faculty)

**Issam Moussallem[1]*, Bilal Hussein[2], Hayssam Kotob[3]**

[1]Pedagogy Faculty, Lebanese University, Beirut, Lebanon
[2] Computer Sciences, university institute of technology, Saida, Lebanon
[3] Pedagogy Faculty, Lebanese University, Beirut, Lebanon
*Corresponding author: issammoussallem@gmail.com

## Abstract

**Purpose –**The purpose of this study is to validate a new measure, the computer competence questionnaire test (Comp2Test), for assessing computer competence.

**Design/methodology/approach –** The study explores three interrelated aspects of the Comp2Test: validity and reliability; instrument dimensionality; and item scales, properties, and qualities within the context of item response theory (IRT), or Rasch modelling. The instrument, for entry-level undergraduates, is based on six dimensions: computer hardware terminology; operating system; Microsoft Word; Microsoft Excel; Microsoft PowerPoint; and Microsoft Internet Explorer and Microsoft Mail. A multidimensional Rasch model was used for item difficulty and competency estimation, based on these six dimensions. Data were collected from 81 undergraduate student teachers at the Lebanese University's Faculty of Pedagogy. To evaluate and improve the test instrument, this study considers the question: how valid and reliable are the corresponding assessment results of the instrument developed (Comp2Test)?

**Findings –** The findings suggest that the questionnaire is a valid and efficient tool for measuring the computer literacy of new student teachers. The overall reliability of the questionnaire was found to be high ($\alpha$=0.84). The tool-factor structure was developed using IRT, and the item–person map analyses provided evidence for the construct validity of the questionnaire.

**Originality/value –** Teachers can use this scale for evaluation of their own teaching and take remedial actions. Mentors and supervisors can use it for diagnostic purposes, and design professional development courses for teachers. Overall, our study outlines an approach for how Rasch modelling may be used to evaluate and improve multiple choice questionnaire instruments in education.

## Keywords:

Computer literacy assessment, Multiple-choice items, Computer competency assessment, Educational evaluation and measurement, Validity assessment instrument, Multidimensional item response theory, Rasch measurement theory

## 1. Introduction

A new computer competency assessment tool (computer competency questionnaire test (Comp2Test)) was developed and tested to measure students' competency, based on six dimensions: computer hardware terminology (HRD); operating system (OS); Microsoft Word (WRD); Microsoft Excel (XLS); Microsoft PowerPoint (PPT); and Microsoft Internet Explorer and Microsoft Mail (IEM).

The Comp2Test consists of 110 items measuring these six dimensions. The sample for this study comprised 80 entry-level undergraduate students from the Pedagogy Faculty at the Lebanese University (PF-LU). The Comp2Test is an assessment of basic computer skills. It measures a student's competency using Internet browsers and common desktop applications such as word processing programs, PowerPoint, email, and the Internet. An analysis of the structure and reliability of the evaluation test practices was performed in the following areas:

- Evaluating the validity of an objective test MCQ measuring computer literacy among entry-level undergraduates within teacher degree programs at the Lebanese University.
- Using item response theory (IRT) and Rasch measurement theory to evaluate dichotomous test result of computer literacy measures.

IRT is an important method of assessing the validity of measurement scales that is underutilized in the field of psychiatry (American Educational Research Association, American Psychological Association, and National Council of Measurement in Education, 2000). IRT describes the relationship between a latent trait (e.g. the construct that the scale proposes to assess), the properties of the items in the scale, and respondents' answers to the individual items.



The purpose of the study is to pilot revisions and establish the validity and reliability of an instrument that will accurately identify gaps in computer skills, competencies, and knowledge in student teachers working, primarily, as educators, or administrators. Gaps thus identified will be filled by developing academic interventions that directly address the derived educational goals necessary to identify the learning outcomes of the course (Saidfudin *et al.*, 2007; Madison Assessment, 2016).

## 2. Methodology

A total of 120 students participated, all registered on an introductory computing course, in their first semester of study at PF-LU. For collecting data from the students, we distributed the survey instrument during the first class of the fall semester of 2016-2017, preceded by a brief explanation of the purpose and the nature of the study. The survey consisted of demographic data that included gender, age, location, years of computer study, educational experience, and the previous school, while tests explored the student's knowledge HRD, OS, WRD, XLS, PPT, and IEM.

## 2.1 Survey

Most questions were developed internally by author consensus. Our survey was constructed to elicit information in six primary domains, relevant to identifying training needs, and system barriers, and to expanding the use of technology in education practice. These domains were: HRD; OS; WRD; XLS; PPT; and IEM. Data were collected and analysed compositely, using the Conquest software package (Wu *et al.*, 2007), which is developed based on the Rasch measurement model.

## 2.2 Participants

The participants in this study were students enrolled in university courses at PF-LU This is a systematic sample, where all students, in the course for teacher training in primary schools, were selected for the study.

## 2.3 Item type selection

All the Comp2Test response options were Microsoft icons. The response options were written as such to ease scoring, to maintain objective scoring, and to minimize test-taker fatigue. Most items followed a typical multiple-choice format, in which an item was followed by several possible response icons, consisting of the correct answer and several distracter icons. The number of alternative responses to each item on the Comp2Test range from three to four.

### 2.4 Statistical analysis

Marginal maximum likelihood (MML) estimation, and expected a posterior (EAP) methods, were implemented, as prescribed by the ConQuest software (Wu *et al.*, 2007).

The MML method for estimating the item parameters, combined with EAP methods, was used to produce the student ability estimates. The joint prior distribution of student abilities was obtained during the MML item parameter estimation process (Wu *et al.*, 2007). First, we present the evidence that supports the application of the six-dimensional model which, in return, supports the six-domain structure of the questionnaire assessment.

### 2.5 Comparison with the unidimensional model

Here the unidimensional Rasch model is nested in the six-dimensional model, meaning that, by applying some constraints (i.e. constraining all the inter-dimensional correlations to 1.0) to the six-dimensional model, a unidimensional model is obtained. The difference in deviance between the multidimensional model and the unidimensional model approximately followed a chi-square distribution, and can be used to provide the index of model fit (Briggs and Wilson, 2003; DeMars, 2004). The difference in deviance statistics of these two models is 930.3, with 9 degrees of freedom, where the degrees of freedom are the difference in the number of parameters estimated in the unidimensional and multidimensional models. The difference in deviance is statistically significant at the $a$=0.001 level. This provides statistical support for the use of a six-dimensional model, along with the theoretical support for the assessment design for the six content topics.

## 2.6 Item–student maps (ISMs)

The distribution of students' abilities, and the difficulty of each item, can also be presented on an ISM. The item difficulty and student ability can be calculated and displayed together. Figure 1 shows the ISM using data from a computer literacy test. The map is split in to two sides. The left side indicates the ability of students, and the right side shows the difficulty of each item. The ability of each student is represented by hash (#) and dot (.), and items are shown by their item number. Item difficulty and student ability values are transformed mathematically, using natural logarithms, into an interval scale whose units of measurement are termed" logits". With a logit scale, differences between values can be quantified, and equal distances on the scale are of equal size (Wu *et al.*, 2007). Higher values on the scale imply both greater item difficulty and greater student ability.

The letters "M", "S" and "T" represent mean, one standard deviation, and two standard deviations of item difficulty and student ability, respectively. The mean of item difficulty is set to 0. Therefore, for example, items 46, 18 and 28 have an item difficulty of 0, 1, and 1, respectively. A student with an ability of 0 logits has a 50% chance of answering items 46, 60, or 69 correctly. The same student has a greater than 50% probability of correctly answering items less difficult, for example items 28 and 62. In addition, the same student has a less than 50% probability of correctly answering more difficult items, such as items 64 and 119.

By looking at the ISM in Figure 1, we can now interpret the properties of the test. First, the student distribution shows that the ability of students is above average, whereas more than half of the items have difficulties below the average. Second, the students on the upper left side have more competency than the items on the lower right side, meaning that the items were easy and unchallenging. Third, most students are located opposite items to which they are well matched, on the upper right, and there are no students on the lower left side. However, items 101, 40, 86, and 29 are too difficult, and beyond the ability of most students.

Overall, in this example, the students are "cleverer" than most of the items. Many items in the lower right-hand quadrant are too easy and should be examined, modified or deleted from the test. Similarly, some items are clearly too difficult. The advantage of Rasch analysis is that it produces a variety of data displays encapsulating both student and item characteristics that enable test developers to improve the psychometric properties of items (American Educational Research Association, American Psychological Association, and National Council of Measurement in Education, 2000). By matching items to student ability, we can improve the authenticity and validity of items, and develop higher quality item banks, useful for future computer-adapted testing.

This paper used the multidimensional random coefficients multinomial logit model (MRCMLM) to examine the construct validity. In the application of MRCMLM, item parameters and student estimates were calibrated to be on the same logit metric, so that, within a single dimension, all model parameter estimates could be compared on the same scale. A Wright Map, a visual

representation of the relative relations between item and person estimates, was also used. Ideally, the item difficulty distribution should cover the span of the student ability distribution, thus providing accurate measures of student proficiency over the whole scale. The information elicited from students will be maximized when the item difficulty level is close to the student ability level. A lack of items in a difficulty range will lead to large errors in ability estimation. In Figure 1, the computer items cover the student ability distributions of the six dimensions quite well, except for the XLS dimension. The ability distribution is more dispersed for XLS, and it is less peaked for the IEM dimension. In both cases, there are sufficient items along the continuum to allow accurate estimates across the whole range of students. For the PPT dimension, even the more difficult items are relatively easy for the top students. This is related to the nature of the PowerPoint domain, in which most items involve basic command. No higher-order thinking is required for these kinds of items.

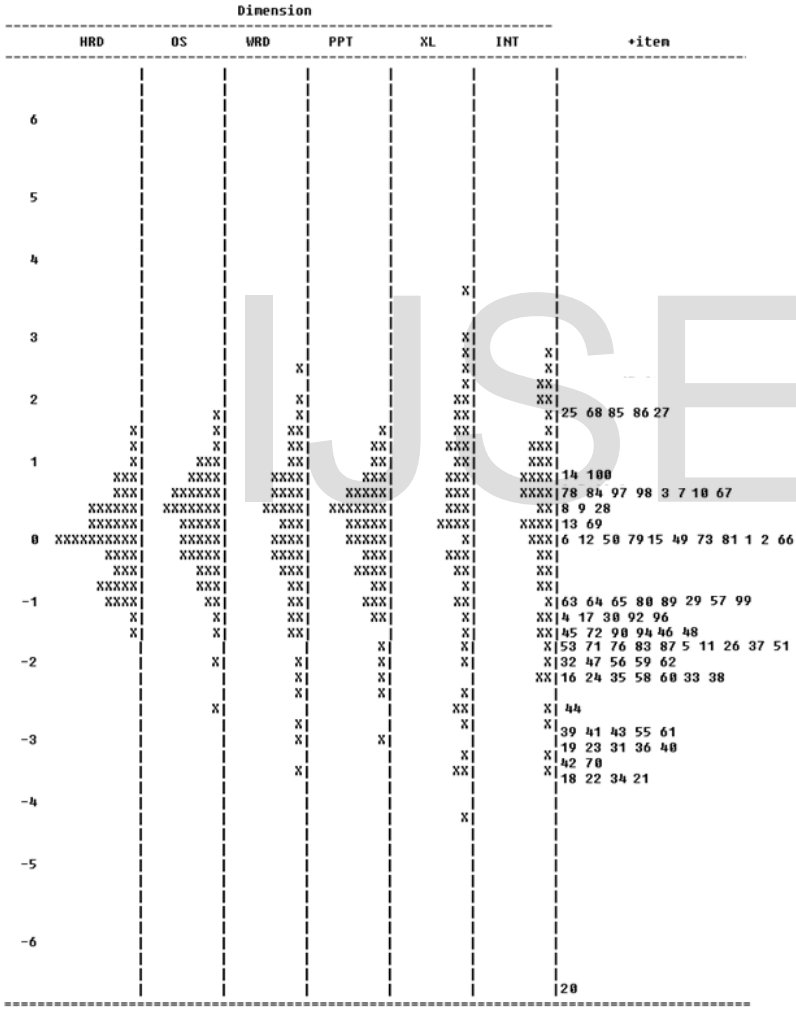


*Figure 1*. Wright Map for the six computer basics dimensions
*Note*: Each "X" represents about 107 cases.

## 3. Findings

### 3.1 Correlation between dimensions

The correlation between dimensions reported from ConQuest ranged from 0.225 to 0.791 (see Table 1). Note that the correlation produced in the ConQuest analysis is not the raw correlation between student ability estimates. These correlations were dis-attenuated or corrected for error, so they are relatively free of measurement noise stemming from various sources (Wu *et al.*, 2007). The OS dimension is highly correlated with the other five dimensions. OS items deal with the basics commands (copy, paste, undo, etc.) and the management of the storage device. These competencies are fundamental skills for carrying out, and completing, other computer activities. A correct answer to items in the other five domains depends, to some extent, on the amount of OS knowledge. The two dimensions that showed the lowest correlation were the IEM and OS dimensions. Except for the fact that they are both "computer basics", these two domains are only loosely connected. On average, the correlations among the six dimensions are reasonably high, given the fact that they all measure student computer competency.



**Table 1: Correlations/covariance between dimensions**

| | | Dimension | | | | | |
|---|---|---|---|---|---|---|---|
| **Code** | Dimension | 1 | 2 | 3 | 4 | 5 | 6 |
| **HRD** | Computer hardware | | 0.156 | 0.173 | 0.137 | 0.430 | 0.449 |
| **OS** | Operating system | 0.300 | | 0.756 | 0.614 | 1.268 | 0.712 |
| **WRD** | Microsoft Word | 0.225 | 0.693 | | 0.953 | 1.788 | 0.668 |
| **PPT** | Microsoft Excel | 0.230 | 0.729 | 0.766 | | 1.262 | 0.969 |
| **XLS** | Microsoft PowerPoint | 0.381 | 0.791 | 0.755 | 0.691 | | 1.836 |
| **IEM** | Internet and mail | 0.450 | 0.503 | 0.319 | 0.600 | 0.597 | |

Note: Multidimensional correlations are shown below the diagonal; consecutive covariance above the diagonal.

The results also demonstrated that violations of the Rasch model assumptions are magnified at higher between-dimension correlations. We recommend that practitioners working with highly correlated multidimensional data use moderate-length (roughly 40 items) instruments, and minimize data-to-model misfit, in the choice of model used for confirmatory factor analysis

(multidimensional random coefficient multinomial logit, or other multidimensional item response theory models).

### 3.2 Reliability

We assessed internal consistency reliability using the separation reliability coefficient. This coefficient is similar to the Cronbach's $\alpha$, except that it uses the metric of person intention scores from the IRM, rather than summed scores across items. We calculated the separation reliability using an expected a posteriori estimation based on plausible values (EAP/PV) scores, rather than the maximum likelihood estimation (MLE) score because a substantial portion of respondents had perfect scores on the Comp2Test. MLE was used only to calculate person parameters, as MLE would exclude perfect scores and would thus underestimate the true reliability of the scale. Again, we used the 0.70 cut-off point for acceptable reliability. The EAP/PV reliabilities of the six dimensions as shown in Table 2.

Only the EAP/PV scale reliability, as a value of internal scale coherence, yielded sufficient values for all four scales.

We also examined the Wright Map to determine if the level of intentions measured by the Comp2Test items covered the full range of person-intention levels. We graphed scale information (inverse of the square of the standard error of measurement) to assess the level of precision at each level of the Comp2Test, again, to determine whether the information was highest at levels where most participants fell.



### 3.3 Reliability coefficients

## Table 2: EAP/PV reliabilities

| Dimension | EAP/PV reliability |
|---|---|
| **HRD** | 0.686 |
| **OS** | 0.820 |
| **WRD** | 0.839 |
| **PPT** | 0.841 |
| **XLS** | 0.889 |
| **IEM** | 0.855 |

The EAP/PV reliability is an estimate for test reliability that is provided by the ConQuest software (Wu *et al.*, 2007), which is obtained by dividing the variance of the individual expected a posteriori ability estimates by the estimated total variance of the latent ability.

The reliabilities obtained from the six-dimensional scaling (EA/PV reliability) are 0.686 for HRD, 0.820 for OS, 0.839 for WRD, 0.841 for PPT, 0.889 for XLS, and 0.855 for IEM,

indicating that when using the correlations between the six scales to draw information from all six of the tests, each test can explain a higher percentage of the variation in students' competency levels.

## 3.4 IRT results

IRT methods (Alagumalai and Curtis, 2010) were used to evaluate the test characteristic curves (TCCs) and test information curves (TICs) of the questionnaire total score and the six cognitive domain scores (Table 3 shows items comprising each domain). A TCC represents a nonlinear regression of the total or domain scores on ability. It can be a very useful tool for evaluating the range of measurement and the degree of discrimination at different points of the ability continuum. In addition, the degree to which the TCC is linear provides an indication of the extent to which the measure provides interval scale or linear measurement (Wu *et al.*, 2007). Furthermore, a TIC relates latent ability to the information (precision of measurement) for the total of domain scores. The information on the x-axis is the reciprocal of the variance of measurement. The TIC provides a means of ascertaining what range of ability levels a test is optimally suited to measure (Baker, 2001).



**Table 3: Items comprising each domain**

| Content domain | Number of items |
|---|---|
| **WRD** | 25 |
| **XLS** | 17 |
| **PPT** | 15 |
| **OS** | 15 |
| **IEM** | 16 |
| **HRD** | 14 |

## 3.5 TCCs

Theoretically, the TCCs model the relationship between an ability level, or theta level, and a raw score for the test. For every level of the ability, the TCC identifies the expected proportion of the raw score to be obtained on the test.

TCCs for the six cognitive domain scores are shown in Figure 2, and the interpretation is shown in Table 4. The TCCs relate latent ability to the expected total domain score (percentage of maximum score) for comparability across domains. All six domain scores showed reduced discrimination at high-ability levels. Item discrimination, in increasing order, is: OS = 1.875 < PPT = 2.364 < IEM = 2.64 < XLS = 3.33 < HRD = 3.889 < WRD = 4.333.

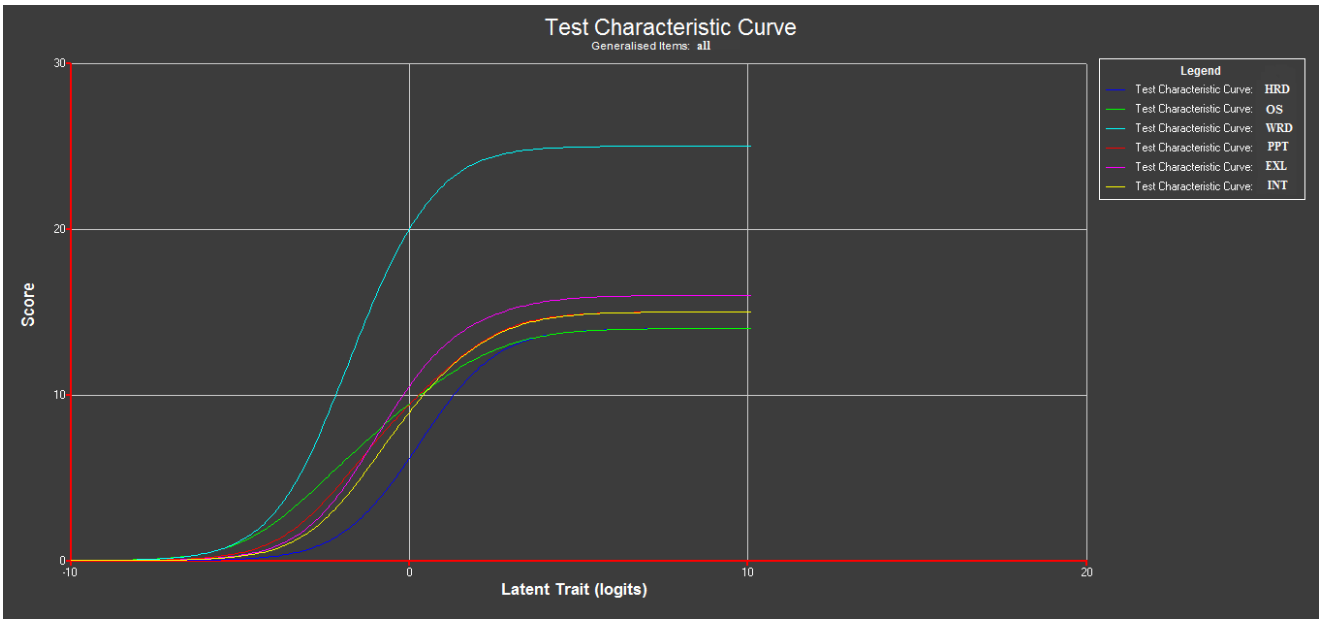


**Figure 2: TCCs for the questionnaire domain scores in the total sample; HRD; OS; WRD; XLS; PPT; and IEM**



**Table 4: Interpretation of the TCCs**

| Module | Test's ability level | Discrimination | Interpretation |
|---|---|---|---|
| HRD | 0.1 logits (corresponds to score of 7=14/2) We expect that students with ability theta = 0.1 on average will answer 7 out of 14 items correctly. | A(0.1,7) B(−1,3.5) Slope=3.889 | • The test's ability level is 0.1 logits, which corresponds to a probability of 0.52. Thus, the test's ability is 52% (students must have an average ability about 52%≈0.1 logits to pass this test). • Item discrimination is 3.889 |
| OS | −1.4 logits | A(−1.4, 7) | • −1.4 logits test ability level |

| | | | |
|---|---|---|---|
| | (corresponds to score of 7=14/2) | B(−3,4) Slope=1.875 | corresponds to a probability of 0.2. Thus, the test's ability level is 20% (students must have an average ability about 20%≈1.4 logits to pass this test).<br>• Item discrimination is 1.875 |
| WRD | −1.8 logits (corresponds to score of 12.5=25/2) | A(−1.8,12.5) B(−3,6) Slope=4.333 | • −1.8 logits test ability corresponds to a probability of 0.14. Thus, the test's ability level is 14% (students must have an average ability about 14%≈1.8 logits to pass this test).<br>• Item discrimination is 4.333 |
| PPT | −0.9 logits (corresponds to score of 7.5=15/2) | A(−0.9,7.5) B−2,4.9) Slope=2.364 | • −0.9 logits test ability corresponds to a probability of 0.29. Thus, the test's ability level is 29% (students must have an average ability about 29%≈0.9 logits to pass this test).<br>• Item discrimination is 2.364 |
| XLS | −0.8 logits (corresponds to score of 8=16/2) | A(−0.8,8) B(1,14) Slope=3.33 | • −0.8 logits test ability corresponds to a probability of 0.31. Thus, the test's ability level is 31% (students must have an average ability about |

| | | | |
|---|---|---|---|
| | | | 31%≈0.8 logits to pass this test).<br>• Item discrimination is 3.33 |
| IEM | −0.5 logits (corresponds to score of 7.5=15/2) | B(2,14.1)<br>A(−0.5,7.5)<br>Slope=2.64 | • −0.5 logits test ability corresponds to a probability of 0.38. Thus, the test's ability level is 38% (students must have an average ability about 38%≈0.5 logits to pass this test).<br>• Item discrimination is 2.64 |

As the slope increases, item discrimination increases. Therefore, as WRD has the highest TCC slope; it has the highest level of discrimination between modules.



### 3.6 TICs

The test information function (TIF) is an extremely useful feature of item response theory. The test information function indicates how well each ability level is being estimated by the test. It basically tells how well the test is doing in estimating ability over the whole range of ability scores. The TIF is simply the sum of all item information functions (IIFs) in the test. While the IIF can tell us the information and precision of a particular item parameter, the TIF can tell us the same thing at the exam level. A TIC relates latent ability to the information (precision of measurement) for the total or domain scores, and it is the summation of the IIFs at each value of theta for all items in the scale. TICs for the six domain scores are shown in Figure 3. The "peak" information of each module is different in each module, where the WRD module has the highest peak (5.2) at theta=−1.8. Therefore, the WRD test provides us with more precise

information. Furthermore, as we have concluded from the TCCs, the discrimination level is highest for the WRD test. Therefore, it discriminates well between students. The level of information provided by each of the six domains varied, with the WRD domain providing the highest precision. The TICs for the WRD domain and the OS domain peaked in the ability range of −5 to 2. For the PPT domain, XLS domain and the OS domain, most test information was contained in the ability range of −3 to 2. The HRD domain provided little information for ability level range of −2 to 3.

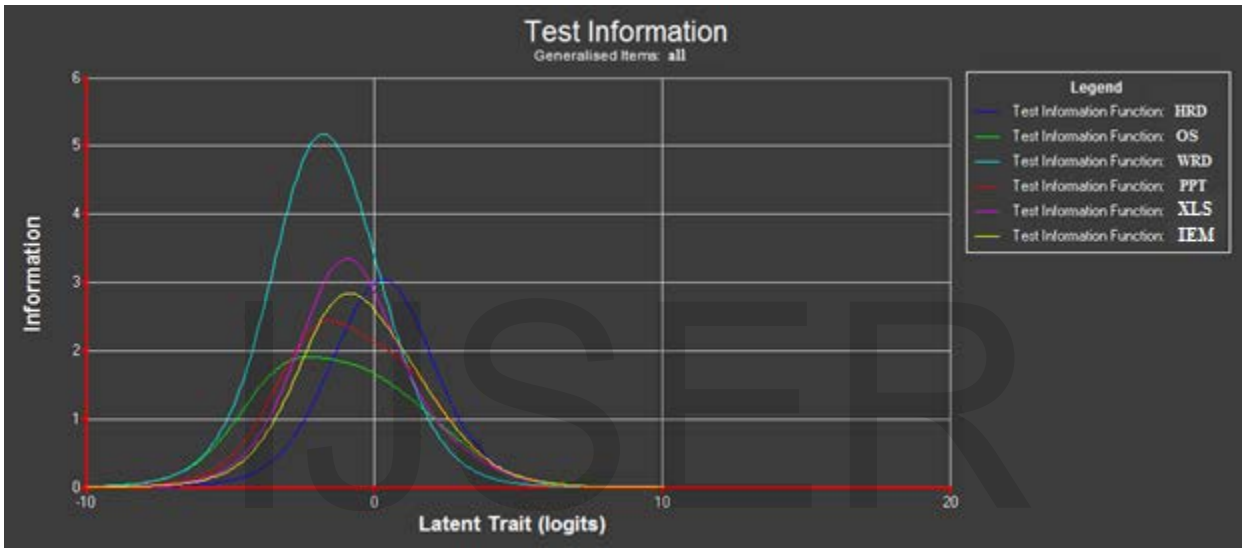


**Figure 3: TICs for the questionnaire domain scores in the total sample: HRD; OS; WRD; XLS; PPT; and IEM**

The TIC of the questionnaire for the total sample peaked at around 0, indicating that the questionnaire mainly provides information for respondents of low to average ability (−3 to 3).

Moreover, all information functions for the different tests reach their maximums for students with an average level of competence. That is to say, close to this zone, it is possible to estimate, precisely, the true level of expertise of the students (information=3, reliability=0.71). Closer to this maximum, however, the accuracy of the test estimate decreases rapidly. Students with a relatively high capacity, which are located in positive space on the graph, reveal a lower, but still sufficient, information value. On the other hand, students with significantly less than average expertise are estimated to have a value of information tending to zero. Thus, the six tests provide many items of measurement related to an average skill level, as well as a few items to measure high levels of proficiency, and also some easy items, designed to measure a low skill level.

## 4. Discussion

IRT analysis suggested that the WRD domain had the best discriminating power and highest precision in measurement for participants in the cognitive function range of ±2SD. The OS domain also demonstrated a high-discrimination ability, although the level of precision is not as high as for the XLS domain. The XLS domain included the items of "IF" formulae, the generation of a pie chart, and selecting the entire Workbook. Our findings supported the questionnaire in screening both XLA and OS dysfunction. These findings can be considered as a guide for educators to strategize their teaching approach and prepare their lessons to focus more, for example, on Excel and Google search operators. Generally, the Rasch measurement model is an effective tool to determine the actual ability of the students, and to diagnose exactly where students are having difficulty the most, in understanding computer basics, and in using statistical formulae in Excel (DeMars, 2004).

Given that students with less than 0.00 logs are classified as "incompetent", they struggle to achieve 56% of the items prescribed in the test. This method measures the ability of students to provide information about the right items that are used in the test. The academician is not only able to measure the competence of students, but also to check the quality of the items. Both provide valuable information for improving learning processes to deliver high-quality education.

## 5. Conclusion

The Rasch model has a role to play in both assessing students through multiple-choice questionnaires, surveys, etc., and in teaching-education research, as a tool for examining the validity and reliability of measures obtained from various test instruments (not just the Comp2Test used in this study) (Astin *et al.*, 2005; Baghaei and Amrahi, 2011). As the example provided in this paper illustrates, the Rasch model may be used to provide an alternative means for measuring student learning ability and can help identify those who may require targeted intervention. Findings from this, and similar, studies may be used to inform future improvements to teaching approaches and styles.

A major strength of this study was the use of Rasch analysis, which allowed a critical psychometric analysis beyond that possible with classical test theory alone.

The Rasch model showed a good fit to the data and confirmed the theoretically modelled levels.

The multidimensional Rasch analysis revealed satisfactory EAP/PV reliabilities, which were between 0.82 and 0.85 for the OS domain and 0.62 for the HRD domain.

## References

Alagumalai, S. and Curtis, D. (2010), "Classical test theory", in Alagumalai, S., Curtis, D. and Hungi, N. (Eds), *Applied Rasch Measurement: A Book of Exemplars*, Springer, Rotterdam, pp. 1-14.

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (2000), *Standards for Educational and Psychological Testing*, American Psychological Association, Washington, DC.

Astin A.W., Banta, T.W., Cross, P.K., El-Khawas, E., Ewell, P.T., Hutchings, P., Marchese, T.J., McClenney, K.M., Mentkowski, M., Miller, M.A., Moran, E.T. and Wright, B.D. (2005), *9 Principles of Good Practice for Assessing Student Learning*, The American Association for Higher Education, Washington, DC.

Baghaei, P. and Amrahi, N. (2011), "Validation of a multiple choice English vocabulary test with the Rasch model", *Journal of Language Teaching and Research*, Vol. 2 No. 5, pp. 1052-1060.

Baker, F. (2001), *The Basics of Item Response Theory*, ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

Bond, T. and Fox, C. (2007), *Applying the Rasch Model*, Lawrence Erlbaum Associates, London.

Briggs, D.C. and Wilson, M. (2003), "An introduction to multidimensional measurement using Rasch models", *Journal of Applied Measurement*, Vol. 4, pp. 87-100.

De Champlain A. (2010), "A primer on classical test theory and item response theory for assessment in medical education", *Medical Education*, Vol. 44, pp. 109-117.

DeMars, C. (2004), "Measuring higher education outcomes with a multidimensional Rasch model", *Journal of Applied Measurement*, Vol. 5, pp. 350-361.

Madison Assessment (2016), *The Information Literacy (ILT) Test Manual*, Madison Assessment, Boulder, CO, available at: https://www.madisonassessment.com/uploads/ILT%20Test%20Manual%20March2016.pdf.

Saidfudin, M., Azlinah M., Azrilah, A.A., Nor Habibah, A. and Sohaimi, Z. (2007), "Appraisal of course learning outcomes using Rasch measurement: A case study in information technology education", *International Journal of Systems Applications, Engineering & Development*, Vol. 4 No. 1, pp. 164-172.

Wu, M.L., Adams, R.J., Wilson, M.R and Haldane, S.A (2007), *ConQuest Version 2: Generalised Item Response Modelling Software*, Australian Council for Educational Research (ACER) Press, Camberwell.