Differential Item Functioning in Online Learning Instrument (EPFun)

Juhaida Abdul Aziz, Maslawati Mohamad, Parilah Mohd Shah, Rosseni Din
Universiti Kebangsaan Malaysia, Bangi, Malaysia
Email: jusyamel@gmail.com

Received 25 November 2015; accepted 26 January 2016; published 29 January 2016

Abstract
Providing young learners with high-quality digital primary education and literacy skills are necessary for them to excel academically. The development of the online Fun Learning Instrument known as English Primary Fun (EPFun) is based on the Piaget Constructivism Development theory. This theory could be applied to young learners as it includes cognitive, behaviourist and second-language acquisition theory as well. The purpose of this study is to detect Differential Item Functioning (DIF) in performances between genders among Malaysian young learners in rural primary schools. The data were obtained from the randomly selected respondents of 106 male and 144 female young learners from the target group. The EPFun instrument consists of four constructs: a) Usefulness (USFN), b) Ease of Use (EOU), c) Ease of Learning (EOL) and d) Satisfaction (SAT). To analyse the responses to the instruments, 40 items overall were made up using 5 levels of coded smileys from the 4 constructs. The data were analysed using SPSS and Winsteps version 3.68.2, a Rasch-based item analysis programme. The findings indicate there is no significant difference in DIF performance between male and female young learners. This conclusion is derived from the t-test for Equality of Means, which revealed the results of the Sig. (2 tailed) for every construct (USFN; Sig = 0.558; EOU; Sig = 0.638; EOL; Sig = 0.628 and SAT; Sig = 0.500) larger than 0.5. The findings also revealed that there were only 7 items detected as DIF. However, there is no significant DIF based on the size range of the logits scales. Female young learners excel in one construct-Satisfaction (SAT). However, male young learners are on a par with female young learners with regard to other constructs. Thus, this instrument is free from DIF and could be used as an indicator to gauge young learners’ literacy skills, especially in using hypermedia reading materials in English.

Keywords
DIF, Web-Based Learning, E-Learning, Fun Learning, Young Learners

1. Introduction

As they advocate a 21st-century learning environment, educators are highly encouraged to promote meaningful learning contexts. Hence, in order to promote meaningful learning contexts, educators should conduct activities which integrate the usage of technology (Juhaida, 2014; Nachmias & Segev, 2003). This is because the new generations are interested in and surrounded by new digital technologies. Some of them already have the relevant knowledge and skills in technologies. This exposure facilitates them in communicating, interacting, and reflecting constructively during their learning process (Nachmias & Segev, 2003; Maslawati et al., 2011; Zahrudin et al., 2011; Juhaida, 2014; Juhaida et al., 2014). The digital environment also promotes fun and authentic learning, which is an essential part of young learners’ lives (Cauffman & MacIntosh, 2006; Rosseni, 2010; Rahamat et al., 2011).

Various attempts have been made by educators to ensure that Web-based Fun learning is fundamental throughout the learning process, especially in reading English materials. Along the way, several studies proposed logical procedures for the study of Differential Item Functioning (DIF) (Cauffman & MacIntosh, 2006; Hanizah et al., 2006; Rosseni et al., 2011). Generally, studies on DIF analysis focus on academic areas. According to Stoneberg Jr. (2004), DIF analysis has been infused into a variety of subjects, namely science, mathematics, English, history, economics, and research studies. However, very few studies have focused on gender bias. Thus, these scholars suggested that studies should be conducted on various aspects related to the item structure and arrangement with the aim of eliminating or reducing any gender bias. To support the issue, Siti Rahayah et al. (2008a; 2008b) revealed some findings on students’ achievement based on gender. Their studies have become the local references from different fields to ensure that DIF analysis on the item functioning is essential to confirm its reliability. In some other related DIF studies, Sheppard et al. (2006) investigated the Hogan Personality Inventory across gender and two racial groups (Caucasian and Black) and have revealed 38.4% (53 out of 138 items) gender-based DIF and 37.7% (52 out of 138 items) race-based DIF. These indicate potential bias for the items displaying DIF more for Caucasians than Blacks. Cauffman & MacIntosh (2006) measured the Massachusetts Youth Screening Instrument by identifying race and gender differential item functioning among juvenile offenders. An item is a basic unit in an instrument. In order to create an item, it is essential to ascertain the stability and equality of all participants. Thus, DIF is used to measure items with different functions in a construct. This can be applied to various demographic groups provided that they are of similar capabilities (Tennant & Pallant, 2007). Numerous educational studies highlighted the diversity of learners’ characteristics, especially gender differences. However, few studies used DIF to provide an explanation of the different performance between genders. Therefore, to proceed locally, this study was embarked upon. In this study, the researchers examined three aspects of EPFun: 1) reliability and validity, 2) the influence of DIF on young learners’ gender and 3) difficulty of EPFun items.

2. Methodology

This study used a survey design. The randomly selected sample consisted of 106 male and 144 female young learners from rural primary schools in Malaysia. The data gathered were analysed using SPSS and Winsteps version 3.68.2, a Rasch-based item analysis programme. The EPFun instrument was developed by a group of researchers (Juhaida, 2014; Juhaida et al., 2014; Rosseni et al., 2011). The EPFun instrument consists of 40 items that examine four constructs: a) Usefulness (10 items), b) Ease of Use (10 items), c) Ease of Learning (10 items) and d) Satisfaction (10 items). Subsequently, in order to measure the data gathered, Item Response Theory (IRT) that employs Rasch measurement analysis was utilized to identify the DIF in the instrument. This step is necessary to ensure the quality of the items. Therefore, the items could be further improved, and thus, gender bias could be avoided or reduced.

3. Results

Problem Solution with Rasch-Based Item Analysis

Table 1 and Table 2 display the group statistics to verify the reliability and validity of the items used in the instrument. They are also used to show if there are any differences in mean between the male and female groups in answering the four constructs using the EPFun survey instrument. Based on both tables, it can be concluded that there is no significant difference in Web-based Fun learning for male and female young learners.
To verify whether gender bias existed in the four constructs of the EPFun instrument, further analysis using Winsteps version 3.68.2, a Rasch-based item analysis programme, was administered. Rasch analysis converts raw data from scores to logits. The logits are compared to a linear model to find its odds of success. The logits are within the range of 0 to 1. It is stated that reliability is measured by the ability of the scale to locate the level of the attribute (Bond & Fox, 2011). The purpose is to ascertain its validity; even though the same constructs are given to other groups of respondents in different environments, the same ability can be produced (Bond & Fox, 2011; Tennant & Pallant, 2007).

Two significant elements of validity are criterion and construct (Linacre, 2006). The criterion-related validity analyses the ability to calculate an outcome. Construct validity observes whether the items used are able to reflect the construct measured. As stated by Linacre (2006), point-measured correlations should be positive. Every item should add a significant approach to the construct (Tennant & Pallant, 2007; Bond & Fox, 2011). The suitable item is calculated by the mean-square residual fit statistics (Bond & Fox, 2011). Fit statistics expected value is 1.0, and ranges from 0 to infinity. Abnormalities represent a lack of fit between the items and the model. The lower values than expected can be interpreted as item redundancy or overlap. Bond and Fox (2011) recommended that item mean square for Infit and Outfit scale (Likert/survey) ranges from 0.6 to 1.4.

Table 3 and Table 4 illustrate the statistical summary of items and person reliability index, respectively. The item reliability index is 0.95, whereas the person reliability index is 0.91. This value is positive, as it is near to 1.0 (Bond & Fox, 2011). Hence, EPFun item repetition prediction is also high if it is being conducted on other groups of respondents with similar capabilities (Bond & Fox, 2011). Item separation index is 4.27. It indicates that EPFun items could be divided into 4 strata. Moreover, this provision connotes that the items are 4 times more distributed from mean square error. The items and person reliability index respectively ≥0.8 is acceptable and strong (Bond & Fox, 2011). In addition, the separation index of ≥2.0 is also acceptable (Bond & Fox, 2011).

Table 5 illustrates a summary of point measure correlation (PTMEA CORR) for 40 items in EPFun. All items show a positive value with the index >0.13. The minimum index is 0.13 for item A4 (Usefulness). The maximum index is 0.84 of item D39 (Satisfaction). The positive value of PTMEA CORR indicates that the measuring items should be carefully constructed. All items in EPFun are to measure 4 constructs. This analysis is the basic step to measure the construct validity used to build and validate the EPFun instrument. PTMEA CORR index
Table 3. Summary Statistics of Item Reliability Index.

<table>
<thead>
<tr>
<th></th>
<th>RAW SCORE</th>
<th>COUNT</th>
<th>MEASURE</th>
<th>MODEL ERROR</th>
<th>INFIT MNSQ</th>
<th>ZSTD</th>
<th>OUTFIT MNSQ</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>1082.30</td>
<td>250.00</td>
<td>0.00</td>
<td>0.10</td>
<td>1.01</td>
<td>−0.40</td>
<td>1.13</td>
<td>0.00</td>
</tr>
<tr>
<td>S.D.</td>
<td>72.20</td>
<td>0.00</td>
<td>0.50</td>
<td>0.01</td>
<td>0.55</td>
<td>4.80</td>
<td>0.69</td>
<td>5.30</td>
</tr>
<tr>
<td>MAX.</td>
<td>1178.00</td>
<td>250.00</td>
<td>1.55</td>
<td>0.14</td>
<td>2.28</td>
<td>9.90</td>
<td>2.56</td>
<td>9.90</td>
</tr>
<tr>
<td>MIN.</td>
<td>821.00</td>
<td>250.00</td>
<td>−0.67</td>
<td>0.06</td>
<td>0.47</td>
<td>−6.30</td>
<td>0.39</td>
<td>−7.10</td>
</tr>
<tr>
<td>REAL RMSE</td>
<td>0.12</td>
<td></td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MODEL RMSE</td>
<td>0.11</td>
<td>ADJ.SD</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Summary statistics of person reliability index.

<table>
<thead>
<tr>
<th></th>
<th>RAW SCORE</th>
<th>COUNT</th>
<th>MEASURE</th>
<th>MODEL ERROR</th>
<th>INFIT MNSQ</th>
<th>ZSTD</th>
<th>OUTFIT MNSQ</th>
<th>ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMMARY OF 250 MEASURED PERSONS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td>181.80</td>
<td>40.00</td>
<td>1.90</td>
<td>0.28</td>
<td>1.23</td>
<td>0.40</td>
<td>1.13</td>
<td>0.00</td>
</tr>
<tr>
<td>S.D.</td>
<td>17.70</td>
<td>0.00</td>
<td>1.14</td>
<td>0.08</td>
<td>0.75</td>
<td>2.50</td>
<td>0.76</td>
<td>2.60</td>
</tr>
<tr>
<td>MAX.</td>
<td>207.00</td>
<td>40.00</td>
<td>4.62</td>
<td>0.61</td>
<td>3.96</td>
<td>6.90</td>
<td>3.94</td>
<td>7.10</td>
</tr>
<tr>
<td>MIN.</td>
<td>113.00</td>
<td>40.00</td>
<td>−0.80</td>
<td>0.16</td>
<td>0.22</td>
<td>−4.10</td>
<td>0.22</td>
<td>−4.70</td>
</tr>
<tr>
<td>REAL RMSE</td>
<td>0.35</td>
<td>ADJ.SD</td>
<td>1.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MODEL RMSE</td>
<td>0.29</td>
<td>ADJ.SD</td>
<td>1.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Point measure correlation (PTMEA CORR) of EPFun constructs.

<table>
<thead>
<tr>
<th>ENTRY NO</th>
<th>RAW SCORE</th>
<th>COUNT</th>
<th>MEASURE</th>
<th>MODEL S.E.</th>
<th>INFIT MNSQ</th>
<th>ZSTD</th>
<th>OUTFIT MNSQ</th>
<th>ZSTD</th>
<th>PT-MEASURE</th>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>821</td>
<td>250.00</td>
<td>1.52</td>
<td>0.08</td>
<td>2.28</td>
<td>9.9</td>
<td>2.37</td>
<td>9.9</td>
<td>0.13</td>
<td>A4</td>
</tr>
<tr>
<td>39</td>
<td>1119</td>
<td>250.00</td>
<td>0.25</td>
<td>0.12</td>
<td>0.47</td>
<td>−6.3</td>
<td>0.39</td>
<td>−7.1</td>
<td>0.84</td>
<td>D39</td>
</tr>
</tbody>
</table>

can be increased if misfitting items are dropped from the cluster item measurement.

Figure 1 shows the numbers of respondents and the difficulty of items capability hierarchy above a logit scale. The results confirm that all items are scattered and point towards the respondents’ capability level. The ranking of respondents with high capability (easy to agree) is above the scale, whilst the ranking of low ability respondents (difficult to agree) is below the scale. The items that are difficult to agree upon are A4 and B13, while difficulty to be measured is 1.58 logit on the top scale. In contrast, the simplest items to agree upon are items C25 and D37 with the measurement of −1.85 logit on the lower scale. The item perceived to be difficult could be answered by respondents with high capability. On the other hand, respondents could answer easy items with both high and low capability (Bond & Fox, 2011; Linacre, 2006). Overlapping items measure different elements with diverse levels of difficulty (Bond & Fox, 2011).

DIF exists when a group of respondents manages to score higher than another group on the same item. In this study, this phenomenon could be observed in Figures 2-5. Hence, to support this finding, item parameters should be similar across the population. As such, to determine if DIF exists in the instrument used, three indicators were used (Bond & Fox, 2011; Tennant & Pallant, 2007), namely:

i) t value of < −2.0 or > 2.0
ii) DIF contrast value of < −0.5 or > 0.5
iii) p (Probability) value < 0.05 or > −0.05

The three indicators were examined thoroughly. An item is considered biased if all three conditions appear
Figure 1. Map of EPFun person-item.
and it should be dropped from the instrument. If the item only meets one of the conditions, it should not be rejected but it should be separated and fixed (Rosseni et al., 2011). With respect to the three indicators given, DIF from EPFun constructs can be determined from Figures 2-5.

Figure 2 reveals a good pattern for both male and female respondents, as there is not much gap or distance between both lines (Blue line 1 and Red line 2). Nevertheless, items A06 to A10 exhibit a small gap between the
two lines. This suggests the items are more difficult for male respondents (Blue line 1) and easier for female respondents (Red line 2). The t-test for Equality of Means revealed the result of the Sig. (2 tailed) as 0.558, which is more than 0.5 as the cut-off point for DIF. The visible distance is very close and not much difference can be measured. Thus, it is safe to conclude that there is no gender bias for items in the first construct (Usefulness).

For the second construct as illustrated in Figure 3, the same phenomenon exists. Both lines do not display any
significant differences within the logit scales. The t-test for Equality of Means revealed the result of the Sig. (2 tailed) as 0.638, which is more than 0.5 for DIF contrast value. Again, the visible distance is very close, and not much difference can be measured except for items B11, B12, B17, B19 and B20. Nevertheless, all ten items in the second construct (Ease of Use) do not show any sign of bias, whereas even the hardest item, B12, appears to be easier for male respondents. Item B17 appears to be easier for female respondents. However, the values indicated in the logit scales are not strong enough to make it a biased item. Hence, the conclusion can be drawn that all items that measure the second construct (Ease of Use) show fairness to both genders.

Similar to the first and second constructs, the third construct (Ease of Learning) in Figure 4 also reveals the same phenomenon. DIF contrast value is 0.628, above 0.5. From the graph, the lines indicate that there is a big gap between male and female respondents. Out of ten items in this construct, six items, namely C21, C23, C26, C28, C29 and C30 exhibit distance between the two lines. This suggests that items C21, C23 and C26 are more difficult for male respondents (Blue line 1) and easier for female respondents (Red line 2). Items C28, C29, C30 reflect otherwise. With respect to the items in EPFun questionnaire, items C21, C23 and C26 are asking one’s opinion on how to use the blog. Asking one’s opinion and expecting an answer that can be neither right nor wrong could be considered as a higher-order thinking skill (HOTs) (Juhaida, 2014). Generally, females are known to be able to exercise HOTs better than males (Din et al., 2011; Prensky, 2001; Prensky, 2007). This explanation can be accepted because males tend to take matters lightly, especially if they do not trigger their interest. Males seldom focus on details or irrelevant issues (Hanizah et al., 2006; Sheppard et al., 2006).

For the last construct (Satisfaction), as can be referred to in Figure 5, seven out of ten items are detected to have a gap between both lines (Blue line 1 and Red line 2). This indicates gender bias. The items D31, D38, D39 and D41 are identified to be easier for male respondents. Items D32, D35 and D37 reveal otherwise. In order to support the issue, the t-test for Equality of Means revealed the result of the Sig. (2 tailed) as 0.5, which is equal to 0.5 for DIF contrast value. The value is not much contradicted to the cut of value of the indicator. Though the items indicate gender bias, researchers were reluctant to drop the items. This is because moderate bias items are considered not problematic, especially when its direction is unsystematic (Tennant & Pallant, 2007). Therefore, the items need to be separated, fixed and then reused in the instrument.

Meanwhile, several astonishing findings were revealed with regards to the issue discussed (Sheppard et al., 2006; Zaharudin et al., 2011). Though gender bias has been detected in some items under the construct (Satisfaction), it could not be the reason for the researchers to drop these items. The findings of this study reveal otherwise, indicating that there is no significant difference between male and female young learners in gaining fun during web-based learning. These findings are parallel with another study, which indicates that item biases do not adversely affect the measurement quality and predictive validity of the overall instrument (Sheppard et al., 2006). Hence, there is no difference with regard to the capability of male and female young learners in Fun learning during web-based learning, as perceived by some scholars (Prensky, 2001; Rahamat et al., 2011). They believe that male young learners are expected to be more powerful in certain aspects, especially physical endurance. Female young learners, on the other hand, are more at ease with cognitive matters, namely mathematic and scientific areas. Female young learners are also capable of achieving good results in the domain of technology. For these reasons, it is imperative that a digital learning environment be developed for both genders. Despite all the issues discussed, DIF analysis indicates that female young learners are prone to possess more non-verbal intelligence than male young learners. Male young learners, on the other hand, are prone to have more verbal intelligence (Hanizah et al., 2006). Each learner has his unique learning style; namely, learning strategies, cognitive levels, types of instruction to which they respond best, and perceptions and attitudes toward the nature of knowledge.

4. Conclusion

It is safe to conclude that EPFun is a valid and reliable instrument, since only 7 out of 40 items were detected to be gender biased. The solution was to simplify and improve the sentence construction. They could be retained and easily understood by the young respondents. Generally, Malaysian young learners are robust in struggling to be proficient in their field (Din et al., 2011). The issue of gender bias could be eliminated or reduced only by the educators’ correct perceptions. Whether an instrument is created or adapted, the process of measuring the reliability and validity of the instrument should be the main focus of any researcher.

New and appropriate teaching approaches hopefully will increase students’ achievement and motivation to learn (El-Bakry et al., 2011; Nachmias & Segev, 2003; Norman et al., 2011; Prensky, 2007; Wheeler, 2011; Wood, 2010). With the awareness that the development of web-based Fun learning is possible within the context...
of the primary educational system in Malaysia, appropriate measures could be taken to attract young learners, regardless of their gender differences toward new multimodal modes of learning. In a nutshell, DIF benchmarking could also be administered in primary schools throughout Malaysia.

References


