CLASSICAL TEST THEORY vs. ITEM RESPONSE THEORY

An evaluation of the theory test in the Swedish driving-license test

Marie Wiberg

EM No 50, 2004

ISSN 1103-2685
Classical Test Theory vs. Item Response Theory

An evaluation of the theory test in the Swedish driving-license test

Marie Wiberg

Abstract

The Swedish driving-license test consists of a theory test and a practical road test. The aim of this paper is to evaluate which Item Response Theory (IRT) model among the one (1PL), two (2PL) and three (3PL) parameter logistic IRT models that is the most suitable to use when evaluating the theory test in the Swedish driving-license test. Further, to compare the chosen IRT model with the indices in Classical Test Theory (CTT). The theory test has 65 multiple-choice items and is criterion-referenced. The evaluation of the models were made by verifying the assumptions that IRT models rely on, examining the expected model features and evaluating how well the models predict actual test results. The overall conclusion from this evaluation is that 3PL model is preferable to use when evaluating the theory test. By comparing the indices from CTT and IRT it was concluded that both give valuable information and should be included in an analysis of the theory test in the Swedish driving-license test.
Introduction

The Swedish National Road Administration, abbreviated here as SNRA, is responsible for the Swedish driving-license examination. The driving-license test consists of two parts: a practical road test and a theory test. In this paper the emphasis is on the theory test. The theory test is a standardized test that is divided into five content areas and is used to decide if the test-takers have sufficient theoretical knowledge to be a safe driver as stated in the curriculum (VVFS, 1996:168). The theory test is a criterion-referenced test (Henriksson, Sundström, & Wiberg, 2004) and consists of 65 multiple-choice items, where each item has 2-6 options, and only one is correct. The test-taker receives one point for each correctly answered item. If the test-taker has a score higher or equal to the cut-off score, 52 (80%), the test-taker passes the test. Each test-taker is also given five try-out items together randomly distributed among the regular items, but the score on those are not included in their final total score. (VVFS, 1999:32).

A test can be studied from different angles and the items in the test can be evaluated according to different theories. Two such theories will be discussed here; Classical Test Theory (CTT) and Item Response Theory (IRT). CTT was originally the leading framework for analyzing and developing standardized tests. Since the beginning of the 1970’s IRT has more or less replaced the role CTT had and is now the major theoretical framework used in this scientific field (Crocker & Algina, 1986; Hambleton & Rogers, 1990; Hambleton, Swaminathan, & Rogers, 1991).

CTT has dominated the area of standardized testing and is based on the assumption that a test-taker has an observed score and a true score. The observed score of a test-taker is usually seen as an estimate of the true scores of that test-taker plus/minus some unobservable measurement error (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). An advantage with CTT is that it relies on weak assumptions and is relatively easy to interpret. However, CTT can be criticized since the true score is not an absolute characteristic of a test-taker since it depends on the content of the test. If there are test-takers with different ability levels a simple or more difficult test would result in different scores. Another criticism is that the items’ difficulty could vary depending on the sample of test-takers that take a specific test. Therefore, it is difficult to compare test-takers’ results between different tests. In the end, good techniques
are needed to correct for errors of measurement (Hambleton, Robin, & Xing, 2000).

IRT was originally developed in order to overcome the problems with CTT. A major part concerning the theoretical work was produced in the 1960’s (Birnbaum, 1968; Lord & Novick, 1968) but the development of IRT continues (van der Linden & Glas, 2000). One of the basic assumptions in IRT is that the latent ability of a test-taker is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of a test-taker can be modeled in different ways depending on the nature of the test (Hambleton et al., 1991). It is common to assume unidimensionality, i.e. that the items in a test measure one single latent ability. According to IRT, test-taker with high ability should have a high probability of answering an item correctly.

Another assumption is that it does not matter which items are used in order to estimate the test-takers’ ability. This assumption makes it possible to compare test-takers’ result despite the fact that they have taken different versions of a test (Hambleton & Swaminathan, 1985). IRT has been the preferred method in standardized testing since the development of computer programs. The computer programs can now perform the complicated calculations that IRT requires (van der Linden & Glas, 2000). There have been studies that compare the indices of CTT and IRT (Bechger, Gunter, Huub, & Béguin, 2003). Other studies have aimed to compare the indices and the applicability of CTT and IRT, see for example how it is use in the Swedish Scholastic Aptitude Test in Stage (2003) or how it can be used in test development (Hambleton & Jones, 1993).

Aim

The aim of this study is to examine which IRT model is the most suitable for use when evaluating the theory test in the Swedish driving-license test. Further, to use this model to compare the indices generated by classical test theory and item response theory.
Method: Sample

A sample of 5404 test-takers who took one of the test versions of the Swedish theory driving license test in January 2004 was used to evaluate the test results. All test-takers answered each of the 65 regular items in the test. Among the test-takers 43.4% were women and 56.4% were men. Their average age was 23.6 years (range = 18-72 years, s = 7.9 years) and 75% of the test-takers were between 18 and 25 years.

Method: Classical test theory

A descriptive analysis was used initially which contained mean and standard deviation of the test score. The reliability was computed with coefficient alpha, defined as

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_X^2}\right),$$

where $n$ is number of items in the test, $\sigma_i^2$ is the variance on item $i$ and $\sigma_X^2$ is the variance on the overall test result. Each item was examined using the proportion who answered the item correctly, $p$-values, and point biserial correlation, $r_{pbis}$. The point biserial correlation is the correlation between the test-takers’ performance on one item compared to the test-takers’ performances on the total test score. Finally, an examination of the 10% of test-takers with the lowest abilities performed on the most difficult items was made in order to give clues on how to model the items in the test (Crocker & Algina, 1986). In this study coefficient alpha, $p$-values and $r_{pbis}$ are used from the CTT.
Method: Item response theory

There are a number of different IRT models. In this study, the three known IRT models for binary response were used; the one (1PL), two (2PL) and three (3PL) parameter logistic IRT model. The IRT model (1PL, 2PL, 3PL) can be defined using the 3PL model formula

\[ P_i(\theta) = c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} , \quad i = 1,2,\ldots,n \]

where \( P_i(\theta) \) is the probability that a given test-taker with ability \( \theta \) answer a random item correctly, \( a_i \) is the item discrimination, \( b_i \) is the item difficulty and \( c_i \) is the pseudo guessing parameter (Hambleton et al., 1991). The 2PL model is obtained when \( c = 0 \). The 1PL model is obtained if \( c = 0 \) and \( a = 1 \).

In order to evaluate which IRT model should be used three criteria, summarized in Hambleton and Swaminathan (1985) are used;

Criterion 1. Verifying the assumptions of the model.
Criterion 2. Expected model features
Criterion 3. Model predictions of actual test results

These three criteria are a summary of the criteria that test evaluators can possibly use. These criteria can be further divided in subcategories. Hambleton, Swaminathan and Rogers (1991) suggest that one should fit more than one model to the data and then compare the models according to the third criterion. The authors also suggested methods for examining the third criterion more closely. However, these suggested methods demand that the evaluator can manipulate the test situation, control the distribution of the test to different groups of test-takers and have access to the test. Since we cannot manipulate the test situation and the driving-license test is classified, only methods that are connected with the outcome of the test have been used. There are a number of possible methods to examine these criteria (see for example (Hambleton & Swaminathan, 1985; Hambleton et al., 1991). In this study the methods described in the following three sections will be used to evaluate the models.
1: Verifying the assumptions of the model

A. Unidimensionality

Unidimensionality refers to the fact that a test should only measure one latent ability in a test. This condition applies to most IRT models. Reckase (1979) suggests that unidimensionality can be investigated through the eigenvalues in a factor analysis. A test is concluded to be unidimensional if when plotting the eigenvalues (from the largest to the smallest) of the inter-item correlation matrix there is one dominant first factor. Another possibility to conclude unidimensionality is to calculate the ratio of the first and second eigenvalues. If the ratio is high, i.e. above a critical value the test is unidimensional. In this study the first method described is used, i.e. observing if there is one dominant first factor.

B. Equal discrimination

Equal discrimination can be verified through examining the correlation between item \( i \) and the total score on the test score, i.e. the point biserial correlation or with the biserial correlation. The standard deviation should be small if there is equal discrimination. If the items are not equally discriminating then it is better to use the 2PL or 3PL model than the 1PL model (Hambleton & Swaminathan, 1985). In this study the notation \( a \) is used for the item discrimination.

C. Possibility of guessing the correct answer

One way to examining if guessing occurs is to examine how test-takers with low abilities answer the most difficult items in the test. Guessing can be disregarded from the model if the test-takers with low ability answer the most difficult items wrongly. If the test-takers with low ability answer the most difficult items correctly to some extent a guessing parameter should be included in the model, i.e. the 3PL model is more appropriate than the 1PL or the 2PL model (Hambleton & Swaminathan, 1985).
2. Expected model features

The second criterion expected model features is of interest no matter which model is used. First, the invariance of the ability parameter estimates needs to be examined. This means that the estimations of the abilities of the examinees $\theta = \theta_1, \theta_2, \ldots, \theta_N$ should not depend on whether or not the items in the test are easy or difficult (Wright, 1968) in Hambleton, Swaminathan & Rogers (1991).

Secondly, the invariance of the item parameter estimates needs to be examined. This means that it should not matter if we estimate the item parameters using different groups in the sample, i.e. groups with low or high abilities. In other words there should be a linear correlation between these estimates and this is most easily examined using scatter plots (Shepard, Camilli, & Williams, 1984).

3. Model predictions of actual test results

The third criterion model prediction of actual test results can be examined by comparing the Item Characteristic Curves (ICC) for each item with each other (Lord, 1970). The third criterion can also be examined using plots of observed and predicted score distributions or chi-square tests can be used (Hambleton & Traub, 1971).

A chi-squared test can be used to examine to what extent the models predict observed data (Hambleton et al., 1991). In this study the likelihood chi-squared test provided by BILOG MG 3.0 was used. This test compares the proportion of correct answers on item $i$ in the ability category $h$ with the expected proportion of correct answers according to the model used. The chi-squared test statistic is defined as follows

$$G_i^2 = 2 \sum_{h=1}^{m} \left[ r_{hi} \ln \frac{r_{hi}}{N_h P_i(\theta)} + (N_h - r_{hi}) \ln \frac{N_h - r_{hi}}{N_h [1 - P_i(\theta)]} \right] \quad i = 1, 2, \ldots, n$$
$$h = 1, 2, \ldots, m$$

where $m$ is the number of ability categories, $r_{hi}$ is the number of observed correct answers on item $i$ for ability category $h$, $N_h$ is the number of test-takers in ability category $h$, $\theta$ is the average ability of test-takers in ability category $h$, $P_i(\theta)$ is the value of the adjusted response function of item $i$ at $\theta$, i.e. the probability that a test-taker with ability
\( \bar{\theta}_i \) will answer item \( i \) correctly. \( G_i^2 \) is chi squared distributed with \( m \) degrees of freedom. If the observed value of \( G_i^2 \) is higher than a critical value the null hypothesis is rejected and it is concluded that the ICC fits the item (Bock & Mislevy, 1990).

The ability \( \theta \) is assumed to be normal distributed with mean 0 and standard deviation 1 and is usually examined with graphics; for example a histogram of the abilities. The test information function can be used to examine which of the three IRT models estimates \( \theta \) best (Hambleton et al., 1991).

Finally, by examining the rank of the test-takers’ ability estimated from each IRT model and compare it with how their rank of abilities are estimated in other IRT models the difference between the models was made clear (Hambleton et al., 1991).

**Estimation methods**

The computer program BILOG MG 3.0 was used to estimate the parameters. Therefore, the item parameters and the abilities are estimated with marginal maximum likelihood estimators. Note that in order to estimate the \( c \)-values. BILOG requires that you enter the highest number of options in an item and uses this as the initial guess when it iterates in order to find the \( c \)-value.
Results: Classical test theory

Using CTT descriptive statistics were obtained about the test. The mean of the test score was 51.44 points with a standard deviation of 6.9 points (range: 13-65 points). Fifty-six percent of the test-takers passed the theory test. The reliability in terms of coefficient alpha was estimated to be 0.82. The \( p \)-values ranged between 0.13 (item 58) and 0.97 (item 38). The point biserial correlation ranged between 0.03 (item 58) and 0.48 (item 36). These two variables are plotted against each other in Figure 1 and are formally connected according to the following formula

\[
r_{pbis} = \frac{\mu_i - \mu_X}{s_X} \sqrt{p/(1-p)},
\]

where \( \mu_i \) is the mean for those who answered the item of interest correctly, \( \mu_X \) is the overall mean on the test, \( s_X \) is the standard deviation on the test for the entire group, and \( p \) is the item difficulty (Crocker & Algina, 1986). In Figure 1 it is obvious that there is quite a large variation in these two variables. In general, most items are close to each other and are in the upper right corner and without the two outliers item 56 and item 58 the range for the \( p \)-values are between 0.47 and 0.97.

![Figure 1. Point biserial correlations plotted against \( p \)-values.](image)

The variation of the point biserial correlations indicates that there is a variation in how well the items discriminate. Item 58 is problematic with a \( p \)-value of 0.03 and together with item 56 with a \( p \)-value 0.26 and \( r_{pbis} \) equal to 0.23. It is important to note that indices other than a high value on the point biserial correlation can be of interest. There are items with
high point biserial correlation that are quite easy, for example items 55 and 56, which have the same point biserial correlation but have a huge different on their item difficulty (0.83 and 0.26, respectively). Also, since the theory test is a criterion-referenced test, an item that is valuable for the content cannot necessarily be excluded from a test because it is too easy (Wiberg, 1999). A more detailed information is presented in Table 1 on the values of the point biserial correlations and the $p$-values. However Figure 1 indicates that there is no specific connection between items that are easy and items that are difficult.

Table 1. The $p$-values and the point biserial correlations ($r_{bis}$) for the 65 items*.

<table>
<thead>
<tr>
<th>Item</th>
<th>p</th>
<th>$r_{bis}$</th>
<th>Item</th>
<th>p</th>
<th>$r_{bis}$</th>
<th>Item</th>
<th>p</th>
<th>$r_{bis}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.88</td>
<td>0.26</td>
<td>23</td>
<td>0.82</td>
<td>0.26</td>
<td>45</td>
<td>0.91</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>0.79</td>
<td>0.28</td>
<td>24</td>
<td>0.95</td>
<td>0.24</td>
<td>46</td>
<td>0.91</td>
<td>0.37</td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.26</td>
<td>25</td>
<td>0.85</td>
<td>0.23</td>
<td>47</td>
<td>0.63</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>0.62</td>
<td>0.33</td>
<td>26</td>
<td>0.82</td>
<td>0.32</td>
<td>48</td>
<td>0.93</td>
<td>0.27</td>
</tr>
<tr>
<td>5</td>
<td>0.81</td>
<td>0.31</td>
<td>27</td>
<td>0.96</td>
<td>0.15</td>
<td>49</td>
<td>0.96</td>
<td>0.27</td>
</tr>
<tr>
<td>6</td>
<td>0.62</td>
<td>0.39</td>
<td>28</td>
<td>0.91</td>
<td>0.35</td>
<td>50</td>
<td>0.70</td>
<td>0.31</td>
</tr>
<tr>
<td>7</td>
<td>0.63</td>
<td>0.44</td>
<td>29</td>
<td>0.87</td>
<td>0.29</td>
<td>51</td>
<td>0.82</td>
<td>0.41</td>
</tr>
<tr>
<td>8</td>
<td>0.95</td>
<td>0.29</td>
<td>30</td>
<td>0.85</td>
<td>0.27</td>
<td>52</td>
<td>0.73</td>
<td>0.32</td>
</tr>
<tr>
<td>9</td>
<td>0.88</td>
<td>0.36</td>
<td>31</td>
<td>0.69</td>
<td>0.41</td>
<td>53</td>
<td>0.53</td>
<td>0.34</td>
</tr>
<tr>
<td>10</td>
<td>0.82</td>
<td>0.31</td>
<td>32</td>
<td>0.90</td>
<td>0.33</td>
<td>54</td>
<td>0.69</td>
<td>0.19</td>
</tr>
<tr>
<td>11</td>
<td>0.81</td>
<td>0.22</td>
<td>33</td>
<td>0.84</td>
<td>0.33</td>
<td>55</td>
<td>0.83</td>
<td>0.23</td>
</tr>
<tr>
<td>12</td>
<td>0.68</td>
<td>0.27</td>
<td>34</td>
<td>0.95</td>
<td>0.25</td>
<td>56</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>13</td>
<td>0.94</td>
<td>0.33</td>
<td>35</td>
<td>0.61</td>
<td>0.24</td>
<td>57</td>
<td>0.84</td>
<td>0.29</td>
</tr>
<tr>
<td>14</td>
<td>0.81</td>
<td>0.36</td>
<td>36</td>
<td>0.78</td>
<td>0.48</td>
<td>58</td>
<td>0.13</td>
<td>0.03</td>
</tr>
<tr>
<td>15</td>
<td>0.96</td>
<td>0.24</td>
<td>37</td>
<td>0.95</td>
<td>0.22</td>
<td>59</td>
<td>0.78</td>
<td>0.13</td>
</tr>
<tr>
<td>16</td>
<td>0.95</td>
<td>0.22</td>
<td>38</td>
<td>0.97</td>
<td>0.18</td>
<td>60</td>
<td>0.84</td>
<td>0.28</td>
</tr>
<tr>
<td>17</td>
<td>0.79</td>
<td>0.35</td>
<td>39</td>
<td>0.73</td>
<td>0.28</td>
<td>61</td>
<td>0.81</td>
<td>0.13</td>
</tr>
<tr>
<td>18</td>
<td>0.86</td>
<td>0.26</td>
<td>40</td>
<td>0.67</td>
<td>0.37</td>
<td>62</td>
<td>0.85</td>
<td>0.32</td>
</tr>
<tr>
<td>19</td>
<td>0.76</td>
<td>0.31</td>
<td>41</td>
<td>0.88</td>
<td>0.41</td>
<td>63</td>
<td>0.59</td>
<td>0.33</td>
</tr>
<tr>
<td>20</td>
<td>0.97</td>
<td>0.20</td>
<td>42</td>
<td>0.87</td>
<td>0.23</td>
<td>64</td>
<td>0.47</td>
<td>0.29</td>
</tr>
<tr>
<td>21</td>
<td>0.94</td>
<td>0.23</td>
<td>43</td>
<td>0.94</td>
<td>0.12</td>
<td>65</td>
<td>0.52</td>
<td>0.23</td>
</tr>
<tr>
<td>22</td>
<td>0.84</td>
<td>0.41</td>
<td>44</td>
<td>0.82</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Items of special interest are in bold type.
Modeling the items

Figure 1 shows no distinct relationship between item discrimination and difficulty. This suggests that both these parameters are important and should be included in a model. There is also of interest to examine whether guessing is present or not. One possible way is to choose test-takers with the 10% lowest ability on the overall test and study how they perform on the more difficult items (Hambleton & Swaminathan, 1985). The five most difficult items were items 53, 56, 58, 64 and 65. Among the test-takers with the 10% lowest ability, 87%, 16%, 11%, 28% and 32% managed to answer these items correctly. These items had each four options except for item 65, which had three options. The values of the observed percentage correct can be compared with the theoretical values if the test-takers are randomly guessing the correct answer, i.e. 25% for the first four items and 33% for item 65. Even though these values are partly close. The overall conclusion is that a guessing parameter should be a part of the model since the observed and the theoretical values are not the same and the test have multiple-choice items where there is always a possibility to guess the correct answer.

Results: Item response theory

1. Verifying the assumptions of the model

In order to verify the first criterion that Hambleton and Swaminathan (1985) used, the assumptions of the three possible IRT models; 1PL, 2PL, and 3PL need to be examined. These three models all rely on three basic assumptions;

1. Unidimensionality
2. Local independence
3. ICC, i.e. each item can be described with an ICC.

The first assumption unidimensionality can be examined using coefficient alpha or most commonly by a factor analysis. The coefficient alpha was 0.82, i.e. a high internal consistence which indicates unidimensionality. The factor analysis gives one distinct factor and many small factors as can be seen in Figure 2. However, it should be noted that the first factor, which has an eigenvalue of 5.9 only accounts for 9 percent of the variance. It would definitely be preferable if more variance was accounted for by the first factor. However Hambleton (2004) explained that this is not
uncommon and as long as there is one factor with distinctively larger eigenvalues it is possible to assume that there is unidimensionality in the test. Note that there are 18 factors with relevant eigenvalues above 1 and together they account for 40% of the total explained variance.

![Figure 2. Eigenvalues from the factor analysis.](image1.png)

![Figure 3. Item discrimination examined using point biserial correlation.](image2.png)

The second assumption local independence was only examined through personal communication with Mattsson (2004) at SNRA. Mattsson assured that no item gives a clue to any other item’s answer. In the future it
would be interesting to review this criterion more carefully. The third assumption simply states that the items can be modeled with an ICC. This is most easily examined by plotting the corresponding ICC for all items in the test (Lord, 1970). If the 2PL IRT model is used the item difficulty is the point where a test-taker has an ability of 0.5 to answer an item correctly, the item discrimination is the slope of the curve at that point. The 3PL model also has a guessing parameter which is the intercept of the curve (Hambleton et al., 1991). Therefore an item with a reasonably high item discrimination and item difficulty looks like item 4 in Figure 4. In the appendix the ICC of all items in the test are shown. From the figures in the appendix it is clear that we can model the items with an ICC. However, there are some problematic items. For example item 58, in the appendix, has low item discrimination, low item difficulty and one can question its role in the test. However, since it is a criterion-referenced test it might be an important item in another way.

![Figure 4. ICC of item 4 modeled with the 3PL model.](image)

Finally, if we add the results from CTT to this information we can conclude that the items do not have equal discrimination and that the test-takers guess the correct answer of an item sometimes. These results lead us to believe that the 3PL model should be preferred instead of the 1PL or the 2PL model.
2. Expected model features

*Invariance of ability estimates*

To examine if we have invariance among the ability estimates the items were divided into easy and difficult items using the item difficulty. Then, the test-takers’ abilities were estimated using these items and plotted against each other in Figure 5.

Figure 5. Ability estimates from easy and difficult items using the 3PL model.

Figure 5 shows that the estimations of the abilities differ whether easy or difficult items are used. This means that this assumption is not quite fulfilled.
Invariance of item parameters estimates

To examine whether there are invariance among the parameter estimates the test-takers were divided into the categories low ability and high ability depending on how they scored on the test (Hambleton & Swaminathan, 1985). The division between low and high ability was made at the cutoff score, i.e. test-takers with scores below 52 were labeled low ability and test-takers with scores equal to or higher than 52 were labeled high ability. The item difficulties were estimated in each group and plotted against each other in Figure 6. If the estimates are invariant the plots should show a straight line.

Figure 6. Item difficulties estimated with the 1PL, 2PL and the 3PL models using test-takers with low and high ability, respectively.

Figure 6 suggests that the item difficulties are quite invariant. The 1PL model has most item parameters which are invariant. But both the 2PL
and the 3PL model have some invariant item difficulties. Figure 7 shows the item discrimination estimated from test-takers with low and high ability, respectively. In the figure one can see that the estimations are not that invariant since the observations are not on a straight line. There are a few more parameters that are invariant for the 2PL than the 3PL model.

Figure 7. Item discrimination estimated with the 2PL and the 3PL models using test-takers with low and high ability, respectively.

Figure 8 shows the item guessing estimated from test-takers with low and high ability, respectively. In Figure 8 the estimation of $c$ is concentrated around 0.17 for test-takers with high ability but is spread between 0.08-0.65 for test-takers with low ability. The overall conclusion is that the estimations are not particularly invariant for the 3PL model.

Figure 8. Item guessing estimated with the 3PL models using test-takers with low and high ability, respectively.
3. Model predictions of actual test results

*Goodness of Fit*

If large samples are used there is a problem with using Goodness of Fit tests since small changes in the empirical data material will lead to rejection of the null hypothesis that the model fits the data (Hambleton & Swaminathan, 1985; Hambleton & Traub, 1973). However, as an initial study the likelihood ratio goodness of fit test was used in order to test whether or not it is reasonable to model the items according to the one, two or three parameter logistic model. When the 1PL model was used 53 items rejected the null hypothesis that the model fitted the data. For the 2PL and the 3PL models the numbers of items were 25 and 16 respectively.

*Normal distributed abilities*

The ability is assumed to be normal distributed with mean 0 and standard deviation 1 and is demonstrated in Figure 9. The line represents the theoretical standard normal distribution and the bars represents the observed distribution. The conclusion from Figure 9 is that the abilities are approximately normal distributed and therefore this assumption is fulfilled.

![Histogram of the test-takers’ abilities with respect to the theoretical standard normal distribution function.](image)

Figure 9. Histogram of the test-takers’ abilities with respect to the theoretical standard normal distribution function.
Test information functions and standard error

How the models manage to describe the items in the test can be examined using Test Information Functions (TIF), and the standard error (S.E). The ideal TIF contains a lot of information on all the test-takers’ abilities and has a low standard error. The TIF and the S.E. are related according to the formula

\[ TIF = \frac{1}{S.E}. \]

Hambleton (2004) suggested that a TIF ≥ 10 is preferable. The curves in Figures 10, 11 and 12 show that the 3PL model gives more information and has a lower standard error than both the 1PL and the 2PL models. Note that the y-scales are different in these three figures.

Figure 10. Test information function and standard error for the 1PL model.

Figure 11. Test information function and standard error for the 2PL model.
Since the theory test in the Swedish driving license test is a licensure test it is important to find as much information around the cut-off score as possible (Birnbaum, 1968; Wiberg, 2003). In this test the cut-off score is on the same level as the ability level – 0.1. The TIF with the highest value on the TIF at the cutoff score is the 3PL, which has 5.9, followed by the 2PL which has 5.8 and the 1PL which has 4.5. This result suggests that the items should be modeled with the 3PL model.

**Rank of the test-takers**

There is a difference between how different models estimate the test-takers’ ability. If we rank the abilities of the test-takers’ for three different models we get the result shown in Figure 13. In Figure 13 the 2PL and the 3PL models give quite consistent estimations of the test-takers’ abilities while the 1PL model ranks the test-takers’ abilities somewhat differently. Note especially that all models rank the low ability test-takers and the high ability test-takers in the same way but the test-takers in the middle are ranked differently if the 1PL model is compared with either the 2PL or the 3PL model.
Figure 13. The rank of the test-takers’ abilities depending on which model the abilities are estimated from compared with the other models.
Comparing CTT and IRT

There have been a number of studies where CTT and IRT have been compared (see for example (Bechger et al., 2003; Nicewander, 1993). The item discrimination parameter in IRT denoted by \( a \), is proportional to the slope of the item characteristic curve at point \( b \) on the ability scale. \( a \) can take the values \( -\infty \leq a < \infty \) but usually items with \( a \)-values less than zero are discarded from a test (Birnbaum, 1968). Further, \( a \) usually has a value between 0 and 2. Figure 14 shows a scatterplot between the \( a \)-values, estimated from the 3PL model and the values of the point biserial correlations.

Theory states that the correlation should be highly positive. The correlation between the \( a \)-values and the values of the point biserial correlation is 0.753, i.e. they are highly correlated and therefore a high correspondence between the two indices.

The item difficulty parameter in IRT denoted by \( b \), can take the values \( -\infty \leq b < \infty \) (Birnbaum, 1968). Usually \( b \) takes values between -2 and 2. The value of \( b \) corresponds to the point on the scale where the probability of a correct answer on an item is 0.5. However, when using the 3PL model the \( b \)-value corresponds to the point where the probability is 0.5(1+\( c \)) of answering an item correctly (Camilli & Shepard, 1994). Figure 15 shows a plot with the estimated \( b \)-values from the 3PL plotted against \( p \)-values. Theoretical there should be a high negative correlation.
between these indices. The correlation is $-0.861$, i.e. highly negative, as expected in theory.

![Figure 15](image-url)  

Figure 15. Estimated $b$-values from the 3PL model plotted against $p$-values.

The $c$-values from IRT can be compared to the inverted number of the 2-6 options for each item. Figure 15 displays this comparison. Note that if an item has only two options, i.e. the test-takers have a theoretical chance of 0.5 of guessing the correct answer, the $c$-value can be as low as 0.15. Therefore, IRT adds valuable information about the test-takers’ choices. As suspected, the test-takers do not choose the options completely at random.

![Figure 15](image-url)  

Figure 15. $c$-values from the 3PL model plotted against the inverted number of options.
Discussion

This study has aimed to describe the theory test in the Swedish driving license test using both classical test theory and item response theory with most weight on the latter. The most important goal has been to compare the 1PL, 2PL and the 3PL models in order to find the model which is most suitable for modeling the items. Further to compare this model results from CTT. The evaluation using classical test theory showed that the internal reliability was high; 0.82. Further, the \( p \)-values of the items ranged from 0.13 to 0.97 and the point biserial correlation from 0.03 to 0.48. One important conclusion was that the item discrimination was not similar between items. Another important result was that if we want to model the items we should include a guessing parameter in the model since some test-takers with low ability tend to guess the correct answer on the most difficult items.

The evaluation using IRT was performed according to three criteria. The first criterion was verifying the model assumptions. The factor analysis in Figure 2 supports the assumption of unidimensionality. There was one dominant factor that explains more of the variation than the other factors. Even though it would be better if the first factor would account for more explained variance this assumption is considered fulfilled. The item discrimination was concluded to be unequal among the items which lead to the conclusion that we should have that parameter in our model. In other words, the 1PL is less suitable than the 2PL or the 3PL models. Guessing should probably also be included in the model since test-takers with low ability still manage to get some of the most difficult items in the test correct. This result suggests that the 3PL model is preferable over the 1PL and the 2PL models.

The second criterion was to what extent the models’ expected features were fulfilled. Figure 5 shows the test-takers’ abilities depending on how they performed on easy or difficult items. Figure 5 suggest that the estimations of \( \theta \) are invariant in all three IRT models.

Figure 6 shows the item difficulties estimated from test-takers with low and high ability, respectively. The conclusion from Figure 6 is that the estimation of \( b \) is invariant for all three models. Further, Figure 7 shows the item discrimination estimated with test-takers with low and high ability for the 2PL model. The pattern in Figure 7 is quite spread out which suggests that there are no invariance. Finally, item guessing for the
3PL model was estimated in Figure 8 using test-takers with low and high ability, respectively. This figure suggests that \(c\) does not have invariant estimates. Since \(a\) and \(c\) does not have invariant estimates a 1PL model which only consists the item difficulty \(b\) should be preferred instead of the 2PL or the 3PL model.

The third criterion was model prediction of test results. Goodness of Fit tests were used to examine how many items had an ICC that fitted the three models. 53 items did not fit the 1PL model, 25 items did not fit the 2PL model and 16 items did not fit the 3PL model. This result suggests that the 3PL model is preferable to the other models. However, as noted earlier it may be dangerous to use these tests in large samples. Figure 9 shows that the \(\theta\) estimations are approximately normal distributed. There are a few more test-takers with low ability than with high ability. The histogram has the same appearance no matter which model is used.

Finally, the three models were compared. The TIF for the three models were plotted in Figures 10, 11 and 12. Note, when comparing these plots, that the 2PL model has the highest maximum information and that it gives most information between \(-3.5\) and \(-1\) on the ability scale. Between 1 and 2 on the ability scale the three models give approximately the same information. This suggests that the 2PL model should be used. However, since the 3PL model gives more information about the ability level that corresponds to the cut-off score, i.e. \(-0.1\) the 3PL model is to prefer. In Figure 13 the estimates of the test-takers' abilities from the test results according to the three different models have been ranked and compared with each other. The conclusion that can be drawn from Figure 13 is that the 2PL and the 3PL models estimate the abilities of the test-takers similar but the 1PL model is only consistent with the other models at the endpoints.

The comparisons of the models have been performed using a large sample of test results. Note that an extended analysis would contain control over the testing situation, reviewing the items in the test etc. None of the models fits the test results perfectly but no model is ever perfect. The 2PL model and the 3PL model are, in general, to be preferred over the 1PL model. The test evaluated is a criterion-referenced test with multiple choice items and the results suggest that guessing should be part of the model. The overall conclusion is that the 3PL model is most suitable to model the items in the theory test in the Swedish driving-license test.
The 3PL model can be used to model the items to a higher level and adds valuable knowledge about the theory test in the Swedish driving-license test.

The last part of this work aimed at comparing the classic test theory and item response theory indices. From this comparison it can be concluded that the estimates are valid for both CTT and IRT. The item discriminations in Figure 14 are positive linear related and the item difficulties in Figure 15 are negatively linear related as the theory states. Figure 15 shows that the values of the guessing parameters are more spread out than if inverted numbers of options are used. This suggests that IRT gives valuable information about a test-taker’s true knowledge. IRT has the advantage that the estimates of the item parameters are independent from the sample that has been used. This advantage is especially useful when reusing a test a number of times. From the ICC it is clear how the items work, and which ability a test-taker has that perform well on each item. The TIF and the standard error give us a measure of the amount of information that is obtained from the test about a test-taker depending on the test-taker’s ability level. Finally, if both CTT and IRT are used when evaluating items, different dimensions of information are obtained since both CTT and IRT add valuable information about the test.

Further research

There are, of course, many different scientific fields related to licensure tests using both CTT and IRT that can be studied in the future. For example, if the evaluators have access to the test situation it can be manipulated. Further, one can examine the effect of speediness, i.e. if the estimated abilities are the same regarding how much time they have to complete the test. If the evaluators have access to the items before the test is given to the test-takers the assumption of local independence can be more carefully examined and it can help understanding why an item has a certain difficulty, discrimination or a certain value of guessing. Other interesting fields concern what kind of studies that can be made using IRT with a theory test in a driving-license test. For example, studying test equating and differential item functioning would definitely be interesting in the theory test in the Swedish driving-license test.
References


Appendix

**Matrix Plot of Item Characteristic Curves**

Figure A. The ICC for all 65 multiple-choice items in the theory test of the Swedish driving license test.