# 运用项目反应理论评估国际问卷指标

## ——以喜爱阅读指标对本地学生的合适度为例

Applying Item Response Theory to Evaluate International Questionnaire Indices
—Analysis of the Fit of Index of Enjoyment of Reading to Local Students as an Example

沈学珺[1]

（1. 上海市教育科学研究院，上海 200032）

【摘要】

本文采用项目反应理论中的拓广分部评分模型从四个方面评估了国际学生评估项目（PISA）的喜爱阅读指标对中国上海学生的合适度，以此为例探讨了如何运用项目反应理论的有效方法来对国际教育测评项目的问卷指标进行本地化研究。研究结果表明该指标的组成项目对上海学生喜爱阅读程度的区分度在可接受范围内，而且备选项的设置较为合理。但是，指标对于喜爱阅读程度低于平均水平的学生有最高的测量精确度，另外还发现有冗余项目。在删除冗余项目后，精简指标与长指标的综合评分高度相关。

【关键词】 项目反应理论；拓广分部评分模型；PISA

# 1 引言

项目反应理论(Item Response Theory, 简称 IRT)相对于 经典测量理论(Classical Test Theory, 简称 CTT）而言是新型的测量理论。它又称潜在特质理论，其核心理念就是被测者的综合特质，例如学生的阅读素养水平、喜爱阅读的程度和家庭社会经济背景等是无法直接观察得到的潜在特质，只能通过分析被测者对于项目（即试题或问题）的反应（即回答）来得到估算。IRT 主要体现的就是被测者对于项目给出特定回答的概率与被测者的潜在特质以及项目特性之间的关系，它包括一系列相关模型。

IRT 在认知测试的应用中已经较为深入和广泛（国际上大型认知测试如 TOEFL, GRE, 以及 PISA 和 TIMSS 的认知部分的设计、分析、计分等环节均以 IRT 作为主要理论依据），同时 IRT 也越来越多地在国内（郭庆科，陈英敏，2005；杨业兵等，2008；朱宁宁，张厚璨，2003）和国外（Brodin, Fors & Laksov, 2010；Edelen & Reeve, 2007；McCormack, Masse, Bulsara, Pikora & Giles-Corti, 2006；Reeve & Fayers, 2005）被应用于编制、分析和评估非认知问卷指标中。国际学生评估项目（Programme for International Student Assessment, 简称 PISA）的非认知问卷的综合指标就是通过 IRT 来设计和计分的。

PISA 是由经合组织（Organization for Economic Co-operation and Development, 简称 OECD）发起和组织一个学生能力国际比较研究，主要测量 15 岁的学生在多大程度上掌握了全面参与社会所需要的终身学习能力。由于 PISA 的测评对象是国际学生样本，其通过 IRT 评估指标的性能中主要考虑的对象是从各国抽取的全体学生样本。在本文中，为了评估国际问卷指标对中国上海本地学生的合适度，所针对的群体仅仅是我国上海市的 15 岁学生，所运用的 IRT 模型是同时提供位置参数和区分度参数的拓广分部评分模型（Generalized Partial Credit Model, 简称 GPCM)。本文以 PISA 2009 中的喜爱阅读指标为例，来说明如何用 IRT 来评估国际问卷指标针对本地学生的性能和提高指标对本地学生的合适度。

# 2 方法

### 2.1 应用 IRT 评估国际问卷指标对本地学生的合适度

根据 CTT，计算指标的综合评分的方法是把对于每个项目的回答值叠加起来，用一个总

通讯作者：沈学珺，E-mail: xuejuns@yeah.net
基金项目：上海市教育科学研究项目（B12142）

和或平均值来代表被测者的潜在特质水平。如果运用这样的叠加法进行运算的话，其前提条件是项目的所有备选项，如莱克特(Likert-scale)问题的“非常不同意”、“不同意”、“同意”和“非常同意”备选项，之间在按照顺序排列后俩俩之间是等距的。然而，在多种 IRT 模型，包括 GPCM 中，并没有要求数据有这样的等距性。其次，根据 CTT 计算的项目性能参数都是依赖于特定的被测者样本，对被测者的潜在特质水平的表达也依赖于特定的项目组。而 IRT 中对项目的性能和被测者的能力水平的表达可以没有那样的依赖性，便于跨被测者样本和跨项目组进行比较。除此之外，如下所述，IRT 还能评估指标组成项目的区分度、内容与被测者匹配度、冗余度以及备选项设置的合理性等。

第一，根据区分度参数来判断项目在潜在特质上对被测者的区分度。如前所述，GPCM 模型有两个主要参数：区分度参数(通常以 a 来表示)和位置参数(通常以 b 来表示)。a 越大，项目区分度越高，越能把具有不同潜质水平的学生区分开来。当 a 在[0.4, 2.5）区间内，项目的区分度是在可接受的范围内的（Chen & Cook, 2009）。CTT 中的项目与指标总分之间的相关度（项总计相关）也是测量项目的区分度的，但是较适用于数据是连续的项目。

第二，根据项目难度系数来判断指标的组成项目的内容是否适合被测学生。根据 CTT，项目的难度和被测者的潜在特质水平是无法在同一个量尺上加以比较的。IRT 则可以将项目的位置参数与学生的潜在特质水平的估计值由同一个量尺来表示。项目的内容最适合具有邻近项目位置参数的潜在特质水平的学生。指标的组成项目的位置参数应该均匀地分布于学生潜在特质水平的区间内。

第三，根据项目特征曲线图来判断指标的组成项目回答选项设置是否合理。项目 特征曲线图就是从图形上显示了被测者对于项目给出某项特定回答的概率（纵轴值）与被测者的潜在特质水平（横轴值）之间的曲线关系。每个项目有一个项目特征曲线图，每个备选项由一条曲线表示。如果在以 logit 单位表示的潜在特质水平[-3,3]全距内，一条曲线始终在其他曲线的下方，则说明这个备选项被选到的可能性总是小于其他选项被选到的可能性，那么这个选项的设置是有问题的。

第四，根据信息函数曲线来判断指标的组成项目是否冗余。与 CTT 相比较，IRT 的一个突出优点就是它对于处于不同潜在特质水平的被测者都提供表示水平估算精确度的信息量，并产生一条信息函数曲线。在 CTT 中通常是通过增加项目的数量来提高指标的信度的，这通常会导致项目在内容上产生冗余的现象。如果两个项目的信息函数曲线重合或几乎重合的话，那么这两个项目对于估计潜在特质水平只提供了相同的信息，删去其中一个项目可以达到精简问卷的目的。

当然，IRT 在取得了技术上的突破的同时也存在被测者样本量大和计算量大的弊端。

### 2.2 PISA 数据和所选取的指标

PISA 采用按照概率与抽样元素的规模大小成比例的抽样法，分为学校抽样和学生抽样两个阶段。参加 PISA 2009 的 5115 名上海样本学生代表了该市所有的在 7 年级及以上就读的 15 岁学生。在本研究的 IRT 分析中每个样本学生都是以他们所代表的学生数加权以后进行计算的。

PISA 2009 中在学生层面共收集了 20 个左右的综合指标，涉及学生的家庭社会经济背景、阅读参与度、学习策略以及学校学习环境和风气等各方面。其中，喜爱阅读指标在组成项目的数量上是最多的，内容是最复杂的，因此将 对于这个指标的分析作为一个例子来说明如何利用 IRT 所提供的有效方法来评估国际问卷指标对本地学生的合适度。如表 2 所示，喜爱阅读指标由莱克特项目组成的，要求学生对 11 个陈述表示同意程度，从“非常不同意”（计分为 1)、“不同意”（2)、“同意”（3）和“非常同意”（4）中选择一项。喜爱阅读指标的第 1、4、6、8 和 9 个项目经过反向计分后，所有项目回答值越大，表示学生喜爱阅读程度越高。

PISA 2009 国际数据库中的上海学生的喜爱阅读指标值与本研究中基于 GPCM 计算的指标值（M = 0.006，SD = 1.10）高度相关，相关系数达到了 0.99。喜爱阅读指标是学生阅读参与度的指标之一, 如 PISA 2009 的测试框架（OECD,2009）所述，阅读参与度主要包括以下四个子维度：

1. 对阅读的兴趣和喜爱----喜爱阅读文学作品和功能性文本，满足求知欲；
2. 自我调控阅读活动----能够自我调控阅读活动、所选择的文本和阅读行为；
3. 阅读的社交性----体现阅读的社交性目标和与别人互动的能力；
4. 广泛而多样的阅读实践----指阅读材料的种类以及对每种材料的阅读量。

根据 4 位教育测评专家的意见，如表 2 所示，喜爱阅读指标的 11 个项目涉及上述前 3 个简称为“兴趣”、“调控”和“社交”的子纬度。

# 3 分析结果

## 3.1 基于 CTT 的分析结果

表 1 显示了喜爱阅读指标通过项目求和后得到的指标总分的平均值、标准差、加权后人数以及表示指标内部一致性的克伦巴赫（Cronbach’s）α 系数。指标的 α 系数为 0.85，说明这个指标都有较高的内部一致性。

表 1　指标总分的基于 CTT 的分析结果

| | 喜爱阅读指标 |
|---|---|
| 加权后学生人数 | 96573 |
| 指标总分均值 | 32.61 |
| 指标总分的标准差 | 5.09 |
| 克伦巴赫系数 | 0.85 |

以下表 2 中显示了上海学生对各个组成项目的回答的平均值和标准差。如表 2 所示，在 11 个项目中，有 8 个项目回答的平均值在 2（“不同意”）和 3（“同意”）之间，且均偏向“同意”。而项目 1、6 和 9 回答的平均值介于 3（“同意”）和 4（“非常同意”）之间。

表2　喜爱阅读指标组成项目分析结果概况

| 喜爱阅读指标组成项目的陈述 | 子纬度 | 平均值 | 标准差 |
|---|---|---|---|
| 1. 我只在不得已时才阅读。（反向计分） | 兴趣 | 3.21 | 0.67 |
| 2. 阅读是我最喜欢的一项爱好。 | 兴趣 | 2.83 | 0.73 |
| 3. 我喜欢与别人谈论书。 | 社交 | 2.72 | 0.72 |
| 4. 我发现很难把书看完。（反向计分） | 调控 | 2.97 | 0.78 |
| 5. 如果收到一本书作为礼物，我会很高兴。 | 兴趣 | 2.77 | 0.74 |
| 6. 对我来说，阅读是浪费时间。（反向计分） | 兴趣 | 3.46 | 0.66 |
| 7. 我喜欢去书店或图书馆。 | 兴趣 | 2.83 | 0.78 |
| 8. 我阅读只是为了找到我需要的信息。（反向计分） | 兴趣 | 2.70 | 0.76 |
| 9. 我无法连续几分钟安静地坐着看书。（反向计分） | 调控 | 3.42 | 0.70 |
| 10. 我喜欢对我读过的书发表自己的看法。 | 社交 | 2.83 | 0.77 |
| 11. 我喜欢与朋友们交换书籍。 | 社交 | 2.87 | 0.81 |

## 3.2 基于 IRT 的分析结果

### 3.2.1 根据项目的 a 参数来判断项目的区分度

该指标各项目的区分度均在可接受的范围内。在通过 PARSCALE 软件得到 GPCM 的项目 a 参数(见表 3)后发现，该指标所有项目的 a 参数分布在[0.43,1.23]区间内，属于[0.4, 2.5）的范围。其中，项目 2 和 6 的 a 参数大于 1.0，最能够区分学生的不同喜爱阅读程度。项目 8、10 和 11 的 a 参数在(0.4, 0.5)之间，区分度最低。其余 6 个项目的 a 参数处于 0.5 和 1.0

之间。

### 3.2.2 根据项目难度参数来判断指标的组成项目的内容是否适合被测学生

如表 3 所示，该指标的各项目的位置参数位于[-3.23, 2.46]的区间内，大多数位置参数为负数，说明从整体来说项目对喜爱阅读程度低于平均水平的学生有最高的测量精确度。尤其，项目 6 和 9 的各位置参数均为负数。

**表3 喜爱阅读指标组成项目的3个子纬度和GPCM参数值**

| 喜爱阅读指标组成项目的陈述 | a | $b_1$ | $b_2$ | $b_3$ |
|---|---|---|---|---|
| 1. 我只在不得已时才阅读。 | 0.94 | -2.62 | -1.78 | 0.60 |
| 2. 阅读是我最喜欢的一项爱好。 | 1.10 | -2.37 | -0.68 | 1.28 |
| 3. 我喜欢与别人谈论书。 | 0.81 | -2.56 | -0.57 | 1.76 |
| 4. 我发现很难把书看完。 | 0.59 | -2.40 | -1.37 | 1.14 |
| 5. 如果收到一本书作为礼物，我会很高兴。 | 0.68 | -2.30 | -0.89 | 1.82 |
| 6. 对我来说，阅读是浪费时间。 | 1.23 | -2.45 | -1.98 | -0.14 |
| 7. 我喜欢去书店或图书馆。 | 0.73 | -2.41 | -0.76 | 1.29 |
| 8. 我阅读只是为了找到我需要的信息。 | 0.47 | -2.49 | -0.94 | 2.46 |
| 9. 我无法连续几分钟安静地坐着看书。 | 0.70 | -2.32 | -2.32 | -0.19 |
| 10. 我喜欢对我读过的书发表自己的看法。 | 0.43 | -3.23 | -1.11 | 1.82 |
| 11. 我喜欢与朋友们交换书籍。 | 0.45 | -2.84 | -1.16 | 1.45 |

### 3.2.3 根据项目特征曲线图来判断指标的组成项目回答选项设置是否合理

查看喜爱阅读指标每个项目的特征曲线图后发现，设置这 4 个备选项对于上海学生来说是比较合理的。在 9 个项目（以图 1 显示的项目 2 为例）的特征曲线图中，[-3,3]全距以内，每个备选项的曲线都在某个潜在特质水平段高于其他所有备选项的曲线。仅在 2 个项目的特征曲线图中，第 1 个备选项（“非常不同意”）的曲线在[-3,3]全距内都在其他备选项曲线之下。第 2 个备选项（“不同意”）的曲线在 1 个项目特征曲线中几乎都在其他备选项曲线的下方。

### 3.2.4 根据信息函数曲线来判断指标的组成项目是否冗余。

在该指标所测量的 2 个子纬度有几乎重合的项目信息函数曲线，表示组成项目有所冗余。如前所述，喜爱阅读指标按照其测量内容可以分成 3 个子纬度，分别是阅读兴趣（兴趣）、自我调控阅读能力（调控）和阅读的社交性（社交）。在图 2、3 和 4 中对每个子纬度各项目的信息函数曲线进行了比较。在“兴趣”子维度中，项目 5 和项目 7 几乎重叠，而且项目 5 始终被项目 7 所覆盖（见图 2），说明在[-3,3]的潜在特质水平全距内项目 5 在估计被测的潜在特质水平上所提供的信息始终低于项目 7。另外，项目 8 所提供的信息在[-3,3]的全距内始终都是最低的。因此，可以考虑删除项目 5 和项目 8。相对于其他两个子维度项目，“社交”子维度的三个项目所提供的信息量较小。在该子维度中，项目 10 和项目 11 的信息函数曲线十分接近，并且项目 10 始终被项目 11 所覆盖（见图 4），说明项目 10 为估算被测的潜在特质水平所提供的信息始终低于项目 11，因此可以考虑删除项目 10。在“调控”子维度中，项目 9 对于喜爱阅读程度较低的学生可以提供较多的评估学生潜在特质水平的信息，而项目 4 对于喜爱阅读程度较高的学生能够提供更多的信息，因此这两个项目的信息函数曲线没有重叠（见图 3）。

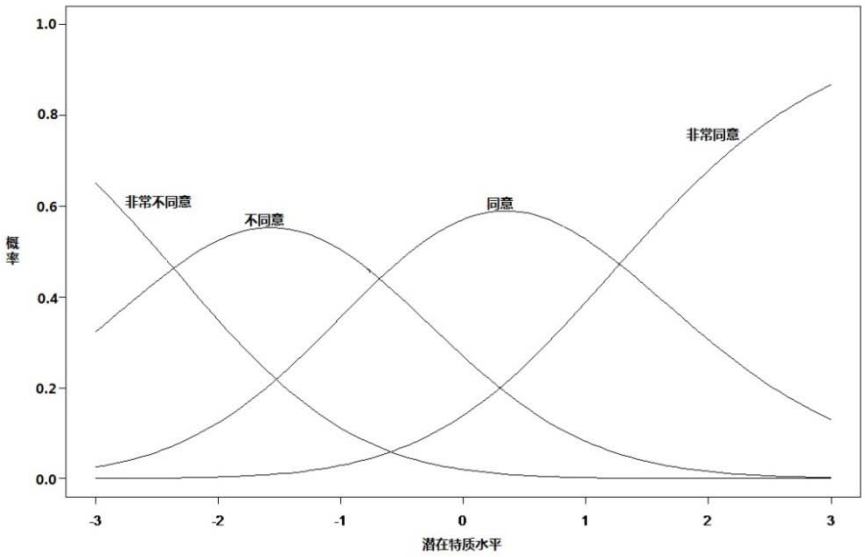


图1　喜爱阅读指标的项目2的项目特征曲线

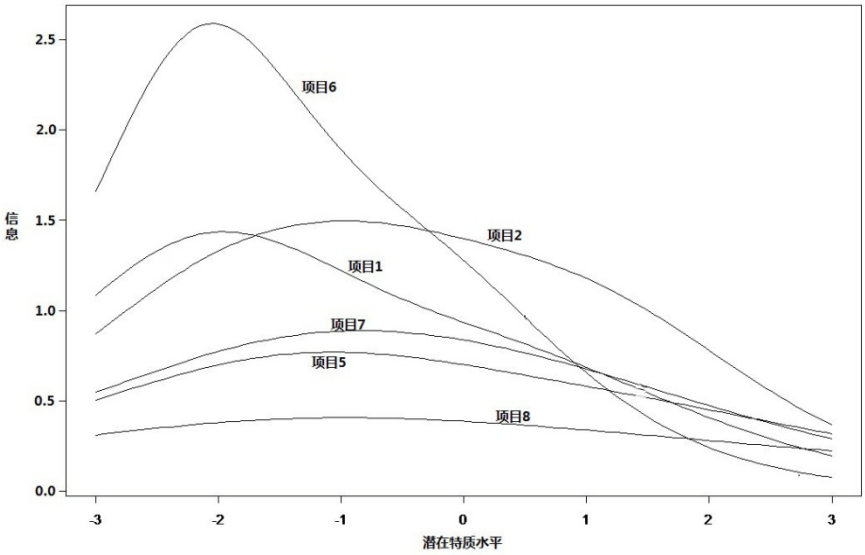


图2　“兴趣”子维度项目的信息函数曲线

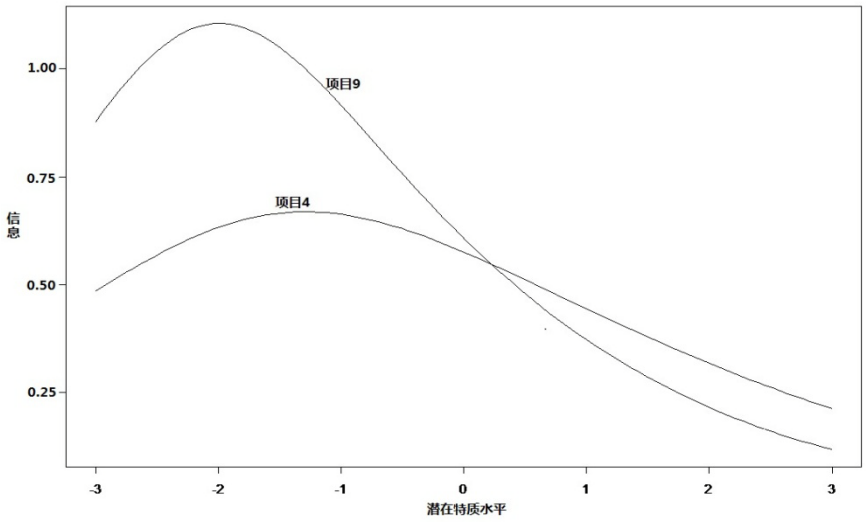


图 3　“调控”子维度项目的信息函数曲线

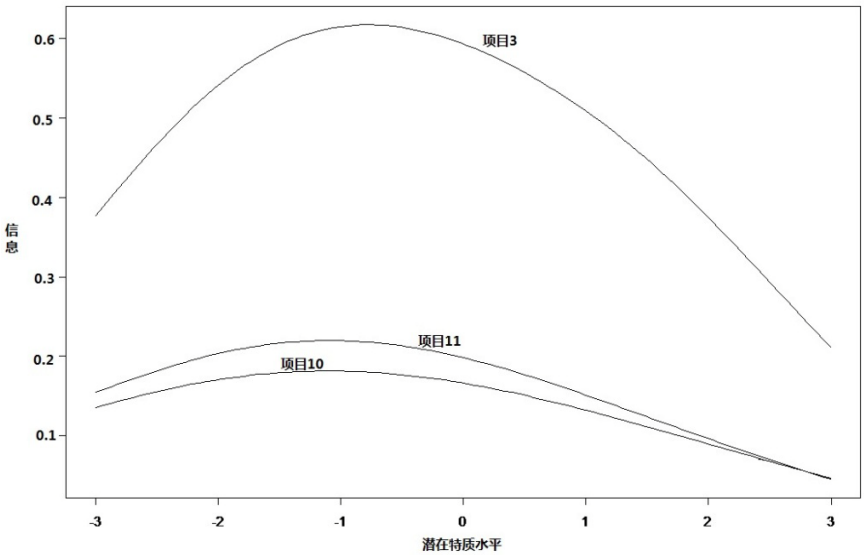


图 4　“社交”子维度项目的信息函数曲线

# 4 讨论

本研究主要运用 IRT 的 GPCM 分析输出结果，从四个方面检测了 PISA 的喜爱阅读指标对本地上海学生的合适度，以此为例子探讨了如何运用 IRT 来进行指标的本地化研究。对项目位置参数的分析表明该指标集中地对喜爱阅读程度低于平均水平的学生有最高的测量精确度。这也体现在项目信息函数曲线图中，这些曲线一般都在负的学生潜在能力值达到峰值，说明对为喜爱阅读程度较低的学生计分提供了最多的信息。在这个情况下，建议对指标进行修订，以使这些指标中有最适合各潜在特质水平的学生的项目。例如，指标中的项目 9 的陈述为“我无法连续几分钟安静地坐着看书”，其各位置参数均为负值，若要提高其“难

度”，可以适当延长陈述中的时间，将其改成“我无法连续十几分钟安静地坐着看书”，以提升其最适合测量的学生的喜爱阅读程度。

对于指标的大部分项目来说，在[-3，3]的潜在特质水平范围内，每个备选项的曲线都在某个潜在特质水平段高于其他所有备选项的曲线。因此对于上海 15 岁学生来说，这个指标设置四个备选项是比较合适的。

在对项目是否冗余的分析中，首先在参考了 PISA 2009 的阅读测评框架后根据项目陈述的文字意思区分了组成项目的三个子维度，然后比较了各个子维度中项目的信息函数曲线。在“兴趣”子维度中，项目 5（陈述为“如果收到一本书作为礼物，我会很高兴”)和项目 7（陈述为“我喜欢去书店或图书馆”）的信息函数曲线接近，并且项目 7 的曲线整体覆盖了了项目 5 的曲线。从项目陈述的文字意义来理解，项目 7 比项目 5 更加直接地询问了学生喜爱书本的程度。另外，项目 8（陈述为“我阅读只是为了找到我需要的信息”）询问学生是否仅仅为了功能性目的去阅读，在该子维度的项目中始终为估算学生喜爱阅读程度提供最低的信息量。这一方面可能由于该项目的陈述表达不够直接，另一方面可能由于为了功能性目的阅读和基于兴趣的阅读并不是那么互相排斥。“社交”子维度的项目 10 和 11 有最小的区分度参数，并且对于估算学生喜爱阅读程度也提供了相似的最小的信息量，项目 10 的信息函数曲线被项目 11 的所覆盖。从项目陈述的意思来理解，项目 10（陈述为“我喜欢对我读过的书发表自己的看法”）和项目 11（陈述为“我喜欢与朋友们交换书籍”）这两个项目都是有关与他人进行有关书籍或阅读内容的交流，相比较之下项目 11 比项目 10 更为具体。

那么，如果根据对项目信息函数曲线的分析结果删除了项目 5、8 和 10，该指标针对上海学生的性能会有怎样的改变呢？这样的话，喜爱阅读指标从 11 个项目的指标精简至 8 个项目的指标，分别被称为长指标和短指标。一般而言，指标的组成项目越多，克伦巴赫 α 系数越高。在减少了了 3 个项目后，指标的 α 系数由长指标的 0.85 略微降低至短指标的 0.83，仅降低了 0.02。在运用 GPCM 计算了每位样本学生的短指标综合评分后发现，长指标的综合评分（M = 0.006，SD = 1.10）与短指标的综合评分（M = 0.008，SD = 1.10）高度相关，达到了 0.98。*t* 检验显示它们的平均分之间并没有显著差异（$t(5113) = 0.96, p = 0.34$）。这说明在略微降低了指标内部一致性的情况下，根据短指标估算的学生喜爱阅读程度与根据长指标估算的基本一致。当然，这里的短指标数据是从所收集的长指标数据里抽取出来的，在这里进行的信度和效度分析仅能为短指标的性能提供参考。

参考文献


Brodin, U., Fors, U. & Laksov, K. B. (2010). *BMC Medical Education, 10*(14). Retrieved March 25, 2013 from http://www.biomedcentral.com/1472-6920/10/14

Chen, S. & Cook, K. F. (2009). SIMPOLYCAT: An SAS program for conducting CAT simulation based on polytomous IRT models. *Behav Res Methods, 41*(2), 499-506.

Edelen, M. O. & Reeve, B. B. (2007). Applying item response theory (IRT) to questionnaire development, evaluation, and refinement. *Qual Life Res, 16*, 5-18.

郭庆科，陈英敏. (2005). 自陈量表式测验应用 IRT 的可行性. *心理学报, 37*(2)，275-279.

Heesch, K. C., Masse, L. C. & Dunn, A. L. (2006). Using Rasch modeling to re-evaluate three scales related to physical activity: Enjoyment, perceived benefits and perceived barriers. *Health Education Research, 21*, 58-72.

McCormack, G. R., Masse, L. C., Bulsara, M., Pikora, T. J. & Giles-Corti, B. (2006). Constructing indices representing supportiveness of the physical environment for walking using the Rasch measurement model. *International Journal of Behavioral Nutrition and Physical*

*Activity, 3(44).* Retrieved March 25, 2013 from http://www.ijbnpa.org/content/3/1/44.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Paris: OECD Publishing.

OECD. (2012). *PISA 2009 technical report*. Paris: OECD Publishing.

Reeve, B. B. & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers and R. D. Hays (Eds.), *Assessing Quality of Life in Clinical Trials: Methods of Practice. 2nd Edition.* Oxford, NY: Oxford University Press. 55-73.

杨业兵，苗丹民，田建全，肖利军，孙菡，洪霞. (2008). 应用项目反应理论对《中国士兵人格问卷》的项目分析. *心理学报, 40*(5), 611-617.

朱宁宁，张厚璨. (2003). CTT与IRT方法对人格测验结果处理的比较研究.*心理学探新, 23,* 48-51.