Advances in Education 教育进展, 2017, 7(4), 177-185
Published Online July 2017 in Hans. http://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2017.74028

Hans汉斯

Using Rasch Model to Analysis Quality of Nine Grade Mathematical Capability Test

Yingying Cai, Jiahao Lin, Dandan Xie, Yifan Zhang, Guangming Li

School of Psychology, Center for Studies of Psychological Application, South China Normal University, Guangzhou Guangdong
Email: 460590758@qq.com

Received: Jun. 23th, 2017; accepted: Jul. 7th, 2017; published: Jul. 12th, 2017

Abstract

In this study, the Rasch model is used to analyze the quality of the nine-year mathematical ability test. We use WINSTEPS software to estimate the parameters of the data, including the difficulty of the item, the level of students' ability and the standard error. Through the Multidimensional testing, the item fitting and the bubble map reflect quality of test. The results show that: 1) the test is basically in line with its measurement objectives. It can better distinguish the level of student ability and the difficulty of item; 2) the item is relatively simple; the level of student ability is greater than the degree of difficulty distribution. This text should add some difficult topics; 3) the existence of individual item and the expected results of the model are inconsistent, needing further analysis of its content and answer status. The nine-year mathematical ability test, which is analyzed in this study, is an accurate test of the quality of the test, which is basically in line with its measurement objectives.

Keywords

Test Quality Analysis, Rasch Model, Student Ability, Item Difficulty, WINSTEPS

基于Rasch模型对九年级数学能力测验进行质量分析

蔡颖颖，林嘉浩，谢丹丹，张一凡，黎光明

华南师范大学心理学院、心理应用研究中心，广东 广州
Email: 460590758@qq.com

收稿日期：2017年6月23日；录用日期：2017年7月7日；发布日期：2017年7月12日

文章引用: 蔡颖颖, 林嘉浩, 谢丹丹, 张一凡, 黎光明. 基于 Rasch 模型对九年级数学能力测验进行质量分析[J]. 教育进展, 2017, 7(4): 177-185. https://doi.org/10.12677/ae.2017.74028

摘　要

使用Rasch模型分析软件WINSTEPS对九年级数学能力测验进行质量分析，通过怀特图(Wright Map)了解该测验的整体情况，包括题目难度，学生能力水平；通过单维性检验、项目拟合度、气泡图等反映题目质量的高低。研究结果表明：① 该测验基本符合其测量目标。能较好地区分出学生能力水平和题目难度。② 题目较简单，学生能力水平范围大于题目难度分布，应增加部分高难度题目。③ 存在个别题目与模型预期的结果不一致，需进一步分析其内容与答题状况。研究所分析的九年级数学能力测验总体上能够准确地进行参数估计，是一套质量较高的测验，该测验基本符合其测量目标。

关键词

测验质量分析，Rasch模型，学生能力，题目难度，WINSTEPS

1. 引言

在我国的教育体系中，老师主要通过考试来了解学生对所学知识的掌握程度，所选用的测验是否能正确反映出学生的真实能力对于提高教学质量非常关键，因此对测验进行质量分析是非常重要的步骤。

我国目前有关测验质量分析的研究主要是基于经典测量理论(CTT)展开的，传统的做法是用测验的总分来反映学生的能力水平(知识掌握水平)，用正确率(或通过率)来定义题目的难度。如果测验的题目很简单，那么学生得分高，意味着能力水平高，相反，如果题目很难，学生得分就低，意味着学生能力低。因此，所反映的学生能力水平的高低是取决于测验的难度，这是测验依赖(test dependent)。同样的道理，在判断题目难度时，如果将题目给能力水平高的学生作答，正确率就高，意味着题目很简单，相反，将同样的题目给能力水平低的学生作答，正确率就低，意味着题目很难，那题目的难易程度，则取决于参加考试的学生能力水平，这是样本依赖(sample-dependent)。在经典测量理论(CTT)中，学生的能力估计和题目的难度估计彼此干扰，并不是客观测量，因此运用经典测量理论进行测验质量分析的结果存在差异。于是研究者们开始关注其他的心理测量理论，薛荣(2007)的研究表明使用项目反应理论(IRT)的概率模型比经典测量理论(CTT)的确定性模型有更多的优势[1]。黄晓婷(2012)、王蕾(2010)和雷新勇(2007)多位学者的研究显示：在项目反应理论中，学生能力水平的估计和题目的估计是独立的，而在大规模考试质量分析中，项目反应理论对题目的分析结果明确，可以使研究者有针对性地采取措施，提高题目的质量[2] [3] [4]。

Rasch 模型是项目反应理论中的单参数模型，由丹麦数学家和统计学家 Georg Rasch 提出的一个潜在特质模型，在 Rasch 模型中，只有一个位置参数来表示题目的难度。根据 Rasch 模型原理，特定的个体对特定的题目做出特定反应的概率可以用个体能力与该题目难度的一个简单函数来表示，个体回答某一题目正确与否完全取决于个体能力和题目难度之间的比较(晏子，2010) [5]。Rasch 模型能克服经典测量理论(CTT)对于测验依赖和样本依赖的问题，所计算出来的难度值与考试的考生平均水平差异在统计学上并没有显著差异，因此运用 Rasch 模型获得题目的难度参数更精确稳定。Rasch 模型所具有的客观性测量，为分析测验的质量提供了一种更为精确的评估方法。

本研究拟运用 Rasch 模型分析软件 WINSTEPS 对九年级数学能力测验进行质量分析，通过怀特图(Wright Map)了解该测验的整体情况，包括题目难度，学生能力水平；通过单维性检验、项目拟合度、气泡图等反映题目质量的高低，为老师以及出题者提高考试命题质量提供测量学参考依据。

2. 研究方法

2.1. 研究资料

本研究所采用九年级数学能力测验题目包含 40 道客观题，在广州市随机选取 1000 名九年级学生完成试卷，取样时尽量涵盖成绩高、中、低学生，对学生的作答情况进行分析。

2.2. 研究工具与统计方法

本研究先使用 SPSS17.0 对测验的数据进行预处理(即进行单维性检验)，随后使用 WINSTEPS3.72 对测验的数据进行 Rasch 分析。

3. Rasch 模型的参数估计结果

3.1. 测验的单维性检验

Rasch 模型是一个理想化的数学模型，它要求所收集的实证数据必须满足事先规定的标准和结构，才能实现客观测量(晏子，2010) [5]。Rasch 模型要求所测量的潜在特质是单维的，也就是说，在本研究中，学生的作答表现只受其所掌握的数学知识影响，没有受阅读理解能力等其他额外变量的影响。因此在利用 WINSTEPS 进行参数估计之前，使用 SPSS 对本研究中的数学能力测验进行探索性因素分析，结果分析如下：

表 1 为 KMO 和 Bartlett 的球形检验，当 KMO 值大于 0.7 及 Bartlett 检验显著性 $P < 0.01$ 时，说明该数据可以进行探索性因素分析。由表可知：KMO 值为 0.97，Bartlett 检验显著性为 $P = 0$，因此可以进行下一步检验。

表 2 是采用主成份分析法提取出来特征根大于 1 的因子，图 1 为因子分析的碎石图。当进行因素分析发现存在多个成分时，若成分 1 与成分 2 的特征根比值超过 5 时，可以说明该数据具有单维性。由表 2 可知，该测验存在多个因子的特征根大于 1(即不只存在一个成分)，成分 1 的特征根为 11.281，成份 2 的特征根为 1.451，成分 1 与成分 2 的特征根比值超过 5，可以说明该测验只受一个因子的影响。

图 1 碎石图中曲线在 X 轴 1 处出现明显的弯折，也可以说明测验只受一个因子的影响。因此可以认为在测验中，学生的作答表现只受其所掌握的数学知识影响，符合 Rasch 模型单维性的要求，可以进行 Rasch 模型分析。

3.2. 题目难度与学生能力变量关系图

图 2 显示了题目难度与学生能力水平间的关系，图中左边代表题目学生能力水平的分布情况，右边

Table 1. The spherical test of KMO and Bartlett
表 1. KMO 和 Bartlett 的球形检验

取样足够度的 Kaiser-Meyer-Olkin 度量		0.97
Bartlett 的球形度检验	近似卡方	11009.52
	df	780
	Sig.	0

Table 2. Explain the total variance
表 2. 解释的总方差

	初始特征值			提取平方和载入		
成份	合计	方差的%	累积%	合计	方差的%	累积%
1	11.281	28.203	28.203	11.281	28.203	28.203
2	1.451	3.627	31.829	1.451	3.627	31.829
3	1.212	3.030	34.860	1.212	3.030	34.860
4	1.083	2.708	37.568	1.083	2.708	37.568
5	1.048	2.620	40.188	1.048	2.620	40.188
6	1.009	2.522	42.710	1.009	2.522	42.710

提取方法：主成份分析。

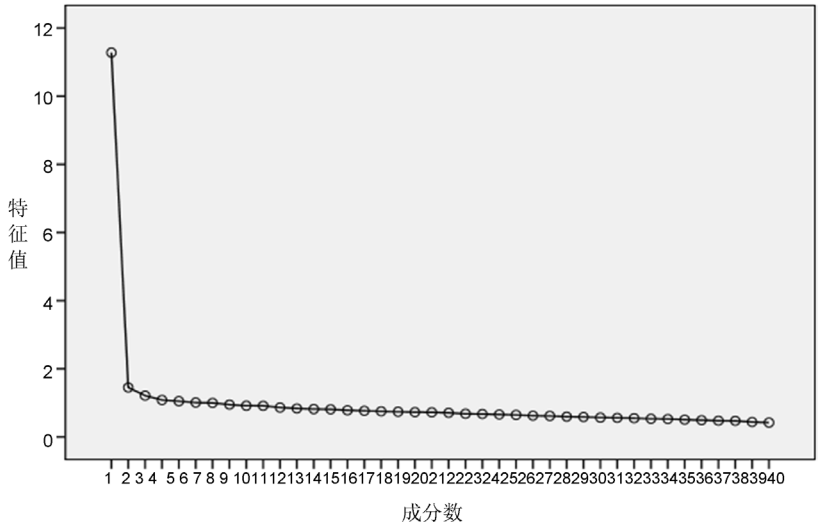

Figure 1. gravel map
图 1. 碎石图

为题目难度的分布。从上往下，学生能力水平逐渐减低，题目难度也逐渐减小。学生之间的距离代表学生能力水平之间的差异，距离靠得越近，差异越小；题目间的距离也如此。处在同一位置的学生能力水平相等，处在同一位置的题目难度相等。当学生能力水平与题目难度越接近时，测验所获得的学生信息量越大，越能精确低估计出学生能力水平。

从图 2 可知，学生能力水平范围宽度约为 5.8 个 logits，分布形式为负偏态；题目难度分布范围约为 4 个 logits，分布形式为正偏态分布。学生能力水平范围大于题目难度分布，题目没有覆盖到 2 到 4 logit 的高能力水平学生。图 2 也清楚地呈现了题目难度的顺序，题目难度基本都在±2 logit 左右，难度分布均匀，其中第 1 题是最简单题目，第 21 题为最难题目。

3.3. 项目拟合和误差统计

表 3 显示的是运用 WINSTEPS 进行参数估计所得到的拟合指标、标准误和相关系数。数据从上到下按照题目的难度进行排序。

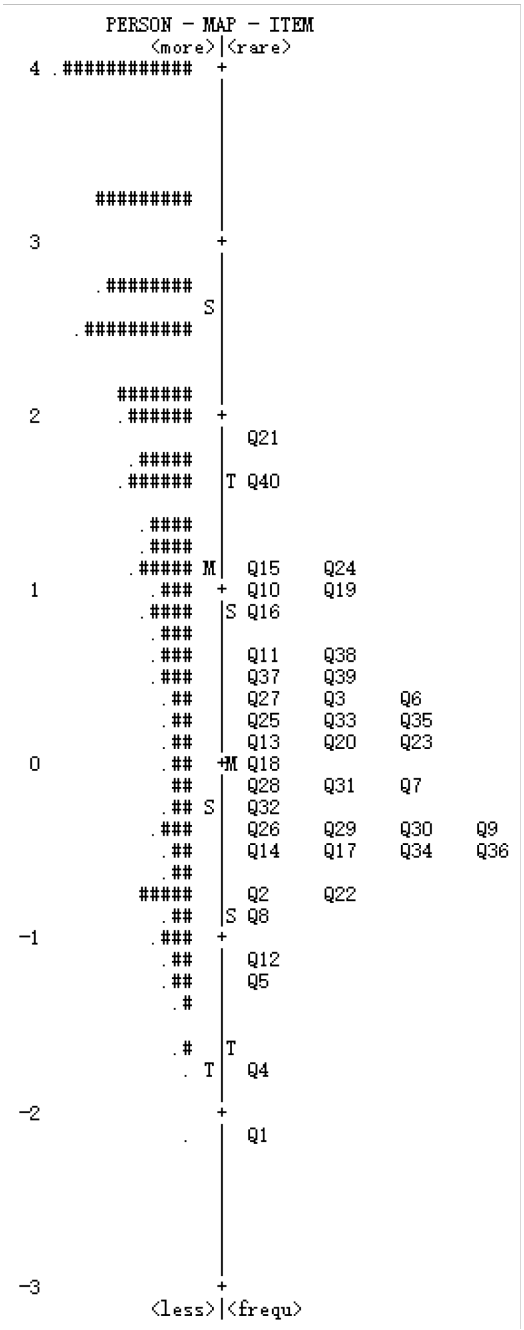

Figure 1. Problem difficulty and student ability diagram (each # mean 10 candidates)
图 1. 题目难度与学生能力关系图(每个#表示 10 名考生)

Rasch 模型分析程序通常报告两个拟合指标：Outfit MNSQ 和 Infit MNSQ，这两个指标都是通过计算残差得到的，Outfit MNSQ 是标准残差的均方，Infit MNSQ 是加权后的残差均方。Outfit MNSQ 和 Infit MNSQ 值为 1 表示数据与模型完全拟合。本研究采用研究者经常使用的 0.7 到 1.3 取值标准来进行拟合度检验。从表 3 可知 Infit MNSQ 的取值范围在 0.80~1.39 之间，题目与模型拟合的很好；Outfit MNSQ 的取值范围在 0.61~1.62 之间，除了第 21、40 题稍微超过了正常的取值范围(0.5~1.5)外，其余 38 题都与模型拟合的很好。

Table 3. Project Fitting and Error Statistics
表 3. 项目拟合和误差统计表

题数	答对人数	Measure	Rasch S.E	Infit MNSQ	Outfit MNSQ	相关系数
21	396	1.92	0.08	1.24	1.52	0.43
40	453	1.57	0.08	1.39	1.62	0.36
15	521	1.16	0.08	1.06	1.16	0.53
24	521	1.16	0.08	0.96	1.01	0.58
10	548	1	0.08	1.31	1.42	0.40
19	549	0.99	0.08	1.02	1.04	0.55
16	577	0.82	0.08	0.93	0.87	0.59
38	609	0.63	0.08	1.01	0.98	0.54
11	614	0.6	0.08	1.14	1.16	0.47
39	622	0.55	0.08	1.23	1.35	0.42
37	628	0.51	0.08	1.03	0.99	0.53
6	640	0.44	0.08	1.14	1.31	0.46
3	655	0.34	0.08	1.1	1.15	0.48
27	659	0.32	0.08	1.07	1.04	0.50
25	665	0.28	0.08	0.85	0.76	0.60
35	671	0.24	0.08	1.05	1.04	0.50
33	676	0.21	0.08	0.82	0.75	0.61
23	694	0.09	0.08	0.9	0.99	0.55
13	695	0.08	0.08	1.07	1.05	0.48
20	697	0.07	0.08	0.83	0.79	0.59
18	701	0.04	0.08	1	0.91	0.52
7	720	−0.09	0.08	0.8	0.71	0.60
31	731	−0.17	0.08	0.91	0.89	0.54
28	734	−0.19	0.08	0.9	0.9	0.54
32	751	−0.31	0.09	0.95	0.83	0.51
30	758	−0.36	0.09	0.91	0.83	0.52
26	763	−0.4	0.09	1.06	1.25	0.43
9	768	−0.44	0.09	1.06	1.09	0.44
29	768	−0.44	0.09	0.88	0.73	0.53
34	773	−0.47	0.09	0.8	0.61	0.57
14	776	−0.5	0.09	0.82	0.73	0.55
17	778	−0.51	0.09	0.92	0.77	0.51
36	779	−0.52	0.09	0.95	0.9	0.49
2	801	−0.7	0.09	0.97	1.13	0.44
22	811	−0.78	0.09	0.89	0.69	0.50
8	823	−0.89	0.09	0.91	0.88	0.47
12	849	−1.13	0.1	0.92	0.82	0.43
5	866	−1.3	0.1	0.9	0.83	0.42
4	905	−1.77	0.12	0.94	0.74	0.36
1	925	−2.06	0.13	1.02	1.28	0.27

标准误是进行数据与 Rasch 模型拟合时的稳定程度，标准误越小说明所得到的结果越稳定。从表 3 中可知，除了第 12、5、4、1 题，其余 36 题的标准误都在 0.10 以下，标准误较小，但 12、5、4、1 题的标准误都在 0.10 以上。

相关系数表示的是题目与题目测量目标的接近程度，相关系数越高，则题目越接近题目的测量目标。除了第一题为 0.27，略低于可接受的最低值为 0.30，其他题目的相关系数都比较高。

3.4. 气泡图

图 3 为气泡图，一个气泡代表一个题目，气泡大小表示 Rasch 标准误，气泡越小说明误差越小，测量的结果就越精确；气泡位置表示题目 Outfit MNSQ 参数大小，气泡越靠近气泡图的中轴线则说明题目与模型拟合得越好。气泡图能形象反映出题目的情况，帮助研究者快速查找出不符合 Rasch 模型的题目。

从图 3 中可以看出，除了题目 21、40 题外，其余题目的 Outfit MNSQ 参数值都在 0.5~1.5 范围内，图中有少部分题目重合在一起，说明题目之间难度水平相近。从图 3 中可以看出，第 21 题难度最大，第 1 题难度最小，且第 1、4 题的气泡大于其他气泡，说明第 1、4 题标准误大，所测量的精确性小，没有准确估计出学生的能力水平。

4. 讨论

4.1. 题目难度与学生能力水平

Rasch 模型通过对数转换，将学生能力和题目难度标定在同一个 logit 量尺上，直接反映题目与题目之间的难度以及学生能力水平和题目难度之间的关系。从题目难度与学生能力变量关系图可知，在九年级数学能力测验中，第 21 题是最难题目，第 1 题为最简单题目。参与测验的学生能力水平远大于题目难度，说明该测验相对于样本学生来讲偏向简单，高能力水平的学生没有相对应难度的题目，应该在测验中增加部分高难度题目，使测验能够对高能力水平的学生进行更精确的估计。

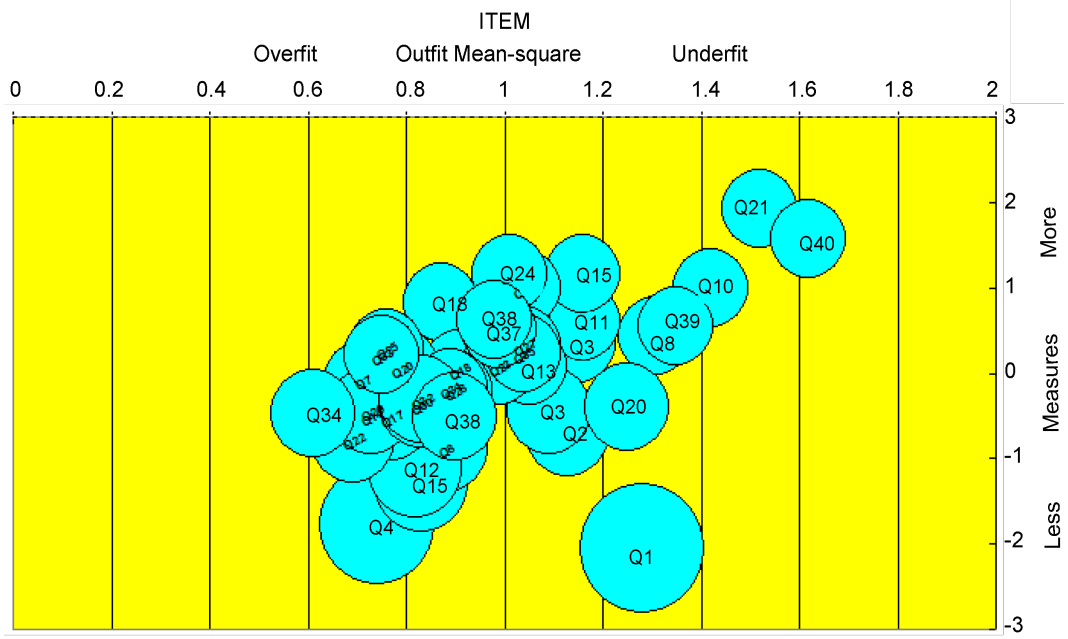

Figure 3. Bubble Chart
图 3. 气泡图

4.2. 测验质量情况

Rasch 模型分析程序 WINSTEPS 所报告的拟合指标 Outfit MNSQ 和 Infit MNSQ，可以反映实证数据与 Rasch 模型的拟合程度，从而可以判断该测验是否测出所要测的潜在特质。当拟合指标的值不好时，可能存在其他影响的无关变量。对九年级数学能力测验进行项目拟合分析发现，Infit MNSQ 的值均处于正常范围，说明学生的能力水平与题目难度比较吻合，两者都能得到比较精确的估计；Outfit MNSQ 的取值，除了第 21、40 题稍微超过了正常的取值范围(0.5~1.5)外，其余 38 题都与模型拟合的很好。这意味着学生在回答该题目时，部分高能力水平的学生选择错误，而低能力水平学生可能由于猜测选择了正确的答案。

除了个别题目，绝多数题目的 Rasch 标准误较小，说明该测验信度比较高，能比较稳定地估计学生的能力水平，题目能较为准确的反映出学生能力。而 Rasch 标准误在 0.10 以上的 12、5、4、1，在结合题目难度与学生能力变量关系图后发现，几题的难度水平较低，其中第 4、1 题都没有相对应的学生能力水平，说明估计学生能力水平时误差较大，需进一步分析其内容与答题状况，确定是否对其进行修改或删除。总体题目的相关系数较高，在可接受的范围水平，表明每道题目与整套样本题目测量的目标一致。

总的来说，该测验大部分题目的拟合度和标准误都处于正常范围内，表明九年级能力测验题目难度有一定的区分度，能较好地反映出学生的能力水平，能较为客观、公平地考查学生的数学知识掌握情况。

4.3. 与模型预期的结果不一致题目的处理

对于那些与模型预期的结果不一致的题目，如第 4、1 题都没有相对应的学生能力水平，且第 1、4 题标准误比较大，说明估计学生能力水平时误差较大，所测量的精确性小，没有准确估计出学生的能力水平，结合所分析的数据可知，这两道题目都比较简单。又如第 21、40 题的 Outfit MNSQ 参数值稍微超过了正常的取值范围(0.5~1.5)，结合所分析的数据可知，这两道题目的难度水平都比较高。

题目与模型预期的结果不一致可能存在很多原因，如题意不清，导致高能力水平的学生钻牛角尖反而答错简单的题目，低能力水平的学生也可能因为使用了某些特殊的解题技巧，因此通过猜测幸运地答对了高难度的题目。而这些题目是否真的需要修正或者删除？不同的研究者有不同的看法。

王文中(2004)认为当发现某个题目不吻合模式预期，这个题目所测量到的特质跟其他题目所测量到的潜在特质并不相同，应该将此一题目排除，但这并不表示该题目不重要[6]。而 Bond 和 Fox(2007)认为拟合度指标可以用来查找表现异常的题目和个体,但它们并不是作为决定是否删除某个题目的简单标准，简单的删除拟合度不好的题目并不是值得提倡的方法[7]。罗德红和龚婧(2015)认为若项目本身符合测量目标的要求，并在学生的最近发展区之间，不存在独立于背景材料的经验性项目等，可以将题目暂时保留下来[8]。因此本研究对于拟合度不好的题目先予以保留，进一步分析其内容与答题状况，看是否在下一次测试中出现同样的情况，若题目本身不符合测量目标要求，则需要对其修改或删除。

5. 结论

本研究使用 Rasch 模型分析软件 WINSTEPS 对九年级数学能力测验进行质量分析。研究结果显示，本研究所分析的九年级数学能力测验总体上能够准确地进行参数估计，大部分题目的拟合程度与标准误差都在接受范围之内，个别题目不符合模型要求。该九年级数学能力测验总体来说是一套质量较高的测验，该测验基本符合其测量目标，测验题目难度有一定的区分度，能较好地反映出学生的能力水平，能较为客观、公平地考查学生的数学知识掌握情况。但是测验的高难度题目较少，略微偏简单，有一部分高能力水平的学生没有被相对应难度的题目所覆盖。应该在测验中增加部分高难度题目，使测验能够对高能力水平的学生进行更精确的估计。而对于那些拟合度不好的题目需进一步分析其内容与答题状况，

根据 Rasch 模型的分析结果可以对题目进行修改，使其能够符合测量的目标。

参考文献 (References)

[1] 薛荣. 从经典测试理论到项目反应理论: 谈语言测试的两种数学模型[J]. 外语研究, 2007(4): 60-64.

[2] 黄晓婷. 从当代教育测量学角度看我国高考研究[J]. 教育与考试, 2012(2): 5-8.

[3] 王蕾. 基于大规模考试的教育质量评价[J]. 教育科学研究, 2010(11): 37-41.

[4] 雷新勇. 用非参数项目反应理论模型研究大规模教育考试维度的问题[J]. 华东师范大学学报教育科学版, 2007, 25(3): 57-64.

[5] 晏子. 心理科学领域内的客观测量——Rasch 模型之特点及发展趋势[J]. 心理科学进展, 2010, 18(8): 1298-1305.

[6] 王文中. Rasch 测量理论与其在教育和心理之应用[J]. 心理测量,2004(27:4): 637-694.

[7] Bond, T.G. and Fox, C.M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Lawrence Erlbaum Associates, Mahwah, NJ.

[8] 罗德红, 龚婧. Rasch 模型在试卷质量分析中的应用——基于五六年级学生阅读素养前测试卷的质量分析[J]. 教育测量与评价, 2015(1): 18-22.