

高级心理测量学丛书

涂冬波◎等著

计算机化自适应测验 理论与方法

JISUANJIHUAZISHIYING
CEYAN

LILUNYUFANGFA



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

高级心理测量学丛书

涂冬波 郑蝉金 戴步云 汪文义◎著

计算机化自适应测验 理论与方法

JISUANJIHUAZISHIYING
CEYAN
LILUNYUFANGFA



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目(CIP)数据

计算机化自适应测验:理论与方法/涂冬波等著. —北京:北京师范大学出版社, 2017. 8
(高级心理测量学丛书)
ISBN 978-7-303-22177-6

I. ①计… II. ①涂… III. ①计算机应用—心理测量学—研究 IV. ①B841.7-39

中国版本图书馆 CIP 数据核字(2017)第 039565 号

营 销 中 心 电 话 010-58805072 58807651
北师大出版社高等教育与学术著作分社 <http://xueda.bnup.com>

出版发行:北京师范大学出版社 www.bnup.com
北京市海淀区新街口外大街 19 号
邮政编码:100875

印 刷:大厂回族自治县正兴印务有限公司
经 销:全国新华书店
开 本:787 mm×1092 mm 1/16
印 张:15.5
字 数:304 千字
版 次:2017 年 8 月第 1 版
印 次:2017 年 8 月第 1 次印刷
定 价:48.00 元

策划编辑:何 琳 责任编辑:齐 琳 马力敏
美术编辑:袁 麟 装帧设计:金基渊
责任校对:陈 民 责任印制:马 洁

版权所有 侵权必究

反盗版、侵权举报电话:010-58800697

北京读者服务部电话:010-58808104

外埠邮购电话:010-58808083

本书如有印装质量问题,请与印制管理部联系调换。

印制管理部电话:010-58805079

总序

心理与教育测量是评价个体心理特质发展水平状态的重要手段。以项目反应理论为代表的现代测量理论的发展,为指导心理与教育测量研究及实践提供了强大的理论与技术支持。在项目反应理论基础上的参数估计、等值、信息量评价、项目功能差异甄别等技术保证了测验开发更加科学。最近十几年蓬勃兴起的认知诊断评价理论,则将测量理论与技术推向了更加精细化的评价水平上。

不过,在国内的心理与教育测量实践中,大多数研究和实践仍然主要是基于经典测量理论基础上的。许多试图使用现代测量理论为指导的研究者由于担心无法很好地把握该理论的原理和方法望而却步。为了让现代测量理论的发展研究成果能够更多地用于指导研究和实践工作,测量学研究者应该做出更多的努力。

江西师范大学心理与教育测量研究中心团队在漆书青、戴海琦、丁树良等教授的带领下,从20世纪80年代初开始对现代测量理论进行深入研究,取得了许多理论和实践研究成果,研究团队也进一步发展壮大。随着研究的深入以及研究领域的进一步拓展,加之现代测量理论受到越来越多研究者的关注,江西师范大学心理与教育测量研究中心团队顺应形势和发展需要,基于自身近30年的理论研究和实践积累,出版一套关于高级心理测量的丛书,这是心理与教育研究领域的一件有益之事,也必将进一步推动心理与教育测量理论与技术在中国的发展。

2012年,该团队曾经出版了一套围绕项目反应理论研究的丛书,该丛书的出版取得了很好的反响。现在,该团队在前期研究和实践的基础上,准备再出版一套关于高级心理测量的丛书,我听后感到非常高兴,且对中国在现代测量理论领域的发展前景充满信心和期待。

高级心理测量学丛书主要有:《计算机化自适应测验:理论与方法》《认知诊断理论》《认知诊断评价理论基础》《高级认知诊断》《Q矩阵理论及认知诊断测验的编制》以及《智慧化测量的理论与技术》等著作。这套丛书包括了当今国际上比较前沿的研究领域,涉及计算机化自适应测验、认知诊断理论和智慧化测量等,这对于推动中国的心理测量学发展及其为实践服务具有重要意义。

值此丛书即将付梓出版之际,作为与江西师范大学心理与教育测量研究团队交流合作多年的同行,我倍感欣慰,特作此短序以示祝贺,并希望他们在今后取得更多的研究成果和更大的发展。

張華華

于美国伊利诺伊大学香槟校区

2017年5月

张华华教授简介

专业领域：心理学、教育心理学、统计学。

美国伊利诺伊大学香槟校区(University of Illinois Urbana-Champaign)终身教授，世界著名心理测量学杂志《应用心理测量》(*Applied Psychological Measurement*)主编，世界心理计量学会 2012—2013 年主席(President of the International Psychometric Society for 2012—2013)，全美教育研究学会院士(AERA Fellow)。

前言

计算机化自适应测验(Computerized Adaptive Testing, CAT)是一种全新的测验形式,采取“因人施测”“量体裁衣”的自适应测量思想,为每个被试/个体选择一份最适合她/他的测验,即为每个被试/个体选择一份具有最小测量误差(最大测量信度)的测量工具,从而真正实现自适应的测量方式。与传统纸笔测验相比,计算机化自适应测验不仅可以达到更高的测量精度,还可以减少测验长度,减轻被试测试负担等优势。计算机化自适应测验也为个体自适应学习(Adaptive Learning)提供了重要的支持。当前关于计算机化自适应测验的研究已受到国内外研究者和应用者的广泛关注。

目前国外已出版了多部关于计算机化自适应测验的学术专著,但我国大陆至今未见相关专著,这不利于计算机化自适应测验在我国的发展,更不利于计算机化自适应测验为我国实践服务,这也是我们撰写此书的由来之一。

本书在作者们多年来对计算机化自适应测验的研究及实践的基础上完成,对计算机化自适应测验的相关理论、方法、技术以及应用等方面进行了较为全面、详细地阐述,为我国在这一领域的发展提供借鉴和参考。

本书写作大纲由涂冬波和郑蝉金两位同志负责,参与本书编写的主要有以下几位同志,在此表示感谢。

第一章:郑蝉金

第二章:涂冬波

第三章:汪文义

第四章:王钰彤 罗照盛

第五章:涂冬波 韩雨婷

第六章:戴步云

第七章:高旭亮 涂冬波

第八章:刘馨婷 涂冬波

第九章:郑屹(美国亚利桑那州立大学)

第十章 第一节:王钰彤

第二节:汪文义

第三节:王潇濛 涂冬波

全书最后由涂冬波同志统稿。该书写作及统稿过程中得到了戴海琦教授、丁树良

教授及美国伊利诺伊大学香槟校区终身教授张华华的帮助，还得到了江西师范大学心理学院各位领导及老师的支持，在此一并表示感谢！

限于时间及能力，本书仍有许多不足之处，恳请广大读者批评指正。

邮箱：tudongbo@aliyun.com

涂冬波

于江西师范大学心理统计与测量中心

2017年5月

目 录

CONTENTS

第一章 计算机化自适应测验 /1

第一节 自适应测验 /2

一、自适应原则在心理测验中的应用 /2

二、自适应测验的主要特征 /4

第二节 计算机化测验 /5

一、划分标准 /6

二、四代计算机化施测系统 /10

第三节 计算机化自适应测验 /15

一、概念 /15

二、CAT 的优势 /17

三、CAT 研究现状与未来研究方向 /19

第二章 计算机化自适应测验理论基础、原理与开发流程 /27

第一节 计算机化自适应测验理论基础：项目反应理论 /28

一、项目反应理论 /28

二、项目反应模型 /28

三、项目反应理论假设 /33

第二节 计算机化自适应测验原理 /36

一、CAT 的产生 /36

二、CAT 的原理 /37

第三节 计算机化自适应测验开发流程 /39

一、准备阶段 /39

二、题库建设 /40

三、CAT 算法确定 /41

四、信度与效度验证 /43

五、正式使用与题库维护更新 /44

第四节 CAT 系统的构成及其测试流程 /45

一、CAT 系统的组成部分 /45

二、CAT 系统的测试流程 /46

第三章 计算机化自适应测验流程中的相关算法与技术 /47

第一节 CAT 选题策略算法 /48

- 一、Fisher 信息量选题方法 /49
- 二、KL 信息量选题方法 /49
- 三、最大优先级指标选题方法 /50

第二节 CAT 参数估计方法 /50

- 一、极大似然估计方法 /50
- 二、贝叶斯估计方法 /51
- 三、加权似然估计方法 /51

第三节 CAT 曝光率控制技术 /52

- 一、条件概率方法 /52
- 二、a 分层方法 /53
- 三、动态 a 分层方法 /53

第四节 CAT 终止策略 /55

- 一、绝对型终止策略 /55
- 二、相对型终止策略 /56

第四章 计算机化多阶段自适应测验 /59

第一节 计算机化多阶段自适应测验简介 /60

- 一、什么是 MST /60
- 二、MST 与题目级别 CAT 的比较 /61

第二节 计算机化多阶段自适应测验设计 /64

- 一、阶段数的确定 /64
- 二、每个阶段中模块数的确定 /64
- 三、每个模块的项目长度 /65
- 四、统计目标与定性规范约束 /65
- 五、计分方法(能力估计) /65
- 六、自适应策略与过程 /65
- 七、组卷 /66

第三节 计算机化多阶段自适应测验的技术与算法 /66

- 一、路由规则 /66
- 二、自动化组卷 /68

第五章 多维计算机化自适应测验 /71

第一节 MCAT 简介 /72

一、多维项目反应理论的发展 /72

二、MIRT 与 FA 的关系 /73

三、MCAT 的产生 /73

第二节 MCAT 中常用 IRT 模型 /74

一、二级评分的多维项目反应模型 /74

二、多级评分的多维项目反应模型 /75

第三节 MCAT 能力估计方法 /76

一、极大似然估计 /76

二、贝叶斯极大后验估计 /78

三、贝叶斯期望后验估计 /79

第四节 MCAT 选题策略 /80

一、基于 Fisher 信息矩阵的选题策略 /80

二、基于 Kullback-Leibler 信息量的选题策略 /82

三、基于互信息的选题策略 /84

四、基于连续香农熵的选题策略 /84

第五节 MCAT 终止规则 /85

一、定长 MCAT 中测验长度的确定方法 /85

二、不定长 MCAT 终止规则 /85

第六章 认知诊断计算机化自适应测验 /87

第一节 CD-CAT 简介 /88

一、认知诊断测验 /88

二、CD-CAT /92

第二节 CD-CAT 中常用认知诊断模型 /93

一、具有认知诊断功能的 IRT 模型 /93

二、规则空间模型 /94

三、统一模型 /95

四、融合模型 /96

五、RRUM 模型 /97

六、DINA 模型 /98

七、高阶 DINA 模型 /98

第三节 认知诊断计算机化自适应测验中的核心技术与算法 /100

一、CD-CAT 的选题方法 /100

- 二、CD-CAT 的参数估计 /107
- 三、CD-CAT 的终止规则 /108
- 四、CD-CAT 的在线标定 /110
- 五、CD-CAT 的总结与展望 /112

第七章 可修改答案的计算机化自适应测验 /115

第一节 可修改答案的计算机化自适应测验简介 /116

- 一、RCAT 优势分析 /116
- 二、RCAT 目前存在的问题 /116

第二节 可修改答案的计算机化自适应测验设计与方法 /118

- 一、基于测试设计视角的 RCAT /118
- 二、基于选题策略视角的 RCAT /123
- 三、基于模型视角的 RCAT /125
- 四、总结与展望 /128

第八章 双因子计算机化自适应测验 /133

第一节 双因子 IRT 模型简介 /134

- 一、基于探索性因素分析模型的研究现状 /135
- 二、基于验证性因素分析模型的研究现状 /138

第二节 BCAT 基本过程 /144

- 一、基于单维视角的 BCAT 过程 /144
- 二、基于多维视角的 BCAT 过程 /148

第三节 BCAT 选题算法 /150

- 一、基于单维的选题策略 /151
- 二、基于多维的最大 Fisher 信息量矩阵选题 /152

第四节 BCAT 在心理学应用中的优势 /153

- 一、拟合性 /153
- 二、既能反映特殊维度情况，也能反映整体情况 /154
- 三、减少被试的作答负担的同时保证测量的准确程度 /154

第九章 CAT 中新题参数在线标定 /157

第一节 概述 /158

- 一、在线标定的概念 /158
- 二、在线标定的基本流程 /159
- 三、在线标定与最优设计 /160

四、在线标定的主要设计因素 /161

第二节 单维二级评分模型下的在线标定 /161

一、试测题选题法 /161

二、在线标定情境下的参数估计方法 /169

三、试测题嵌入位置 /172

四、试测终止规则 /173

第三节 其他模型下的在线标定 /174

一、认知诊断考试的在线标定 /174

二、多维 IRT 的在线标定 /174

三、单维多项评分模型的在线标定 /175

第十章 CAT 技术在实践中的应用 /177

第一节 Multistage-CAT 在 GRE 中的应用 /178

一、为什么选用 MST /178

二、GRE 测验的 MST 设计 /179

三、GRE 测验规范 /180

四、GRE 测验 MST 设计的评估 /180

第二节 CD-CAT 在英语听力理解中的应用 /182

一、题库建设 /182

二、选题策略 /184

三、参数估计 /185

四、真实题库下模拟实验 /185

五、真实测试效度验证 /186

第三节 CAT 在心理健康评估中的应用 /187

一、研究目的 /187

二、研究方法与过程 /188

三、研究结果 /190

四、讨论 /202

五、展望 /203

参考文献 /205

第一章 计算机化自适应测验

与传统的纸笔测验(Paper and Pencil Test, P&P)相比,计算机化适应测验采用“因人施测”“千人千卷”的自适应测量思想。它具有高效、快捷、测量误差小等优势,深受研究者与实践者的推崇。本章重点介绍了自适应测验、计算机化测验以及计算机化自适应测验的产生背景、概念、形成过程、特征及相关算法等,并展望了计算机化自适应测验未来的研究发展方向。

第一节 自适应测验

自适应测验(Adaptive Testing)是一种智能化测验形式(Smart Testing),它能够根据考生已有的作答对其进行能力估计,然后为考生选择最合适的考题。文献中有很多不同的专有名词用来描述这种测验形式,包括定制式测验(Tailored Testing)、个性化测验(Individualized Testing)、程序化测验(Programmed Testing)、序列项目测验(Sequential Item Testing)、作答权变测验(Response-contingent Testing)、计算机化测验(Computerized Testing)以及树状分支测验(Branched Testing)。这些术语反映了这种智能化测验的不同侧面。例如,定制式测验体现了智能化测验为每一位考生“量体裁衣”,提供独特的考试内容;程序化测验说明这个测验是受电脑程序控制的;序列项目测验揭示了智能化测验中——按序选定考题展示给考生的现象;作答权变测验指出了考题的选择是基于已有作答的基本事实;树状分支测验体现了由于每个考生可能在每个节点(考题)被分流到不同的路径(接受不同的考题),因而形成了一个有趣的树状分支图。虽然存在很多不同的载体,但是计算机是最理想最普遍的载体,最能体现这种智能化测验的特点。虽然有很多不同的名字,但是它们都体现了根据考生的表现智能化地开展测验的特点,因此目前被学术界与工业界普遍接受的术语是自适应考试,而主要的施测方式是计算机化自适应测验(Computerized Adaptive Testing, CAT)。

一、自适应原则在心理测验中的应用

早在计算机科学诞生之前,心理学家就把这种自适应的智能化原则运用到心理测验中去了。著名的例子是智力测验中的比内智力测验。早在1905年(请注意在这个时期心理测验还处于早期发展阶段,传统意义中的标准化纸笔测验还没有诞生),比内智力量表就很好地体现了根据考生具体能力调整考题的自适应变化原则。在这个量表中,根据考题难度由浅入深排列,以通过题数的多少作为鉴别智力高低的标准,并且据此提出了智力年龄的概念。比内智力量表的实施过程如下,很好地体现了这个自适应原则。

第一,比内智力测验有一个标定好的题库。比内智力测验中的项目按照难度从低到高排列,并且按年龄水平分组,包括3~11岁。每个年龄组的儿童在解答本年龄组的项目时正确作答的概率大约是50%。

第二,比内智力测验由训练有素的心理学家与考生进行一对一的施测,目的是寻找与每个考生最匹配的难度水平(智力年龄)。这个过程很像跳高运动员的比赛过程。

第三，每个考生有不同的测验起点。比内智力测验开始时，施测的考官需要对考生的能力进行估计，一般都用生理年龄，但是如果有更有效的信息，可以进行调整。

第四，它有一个事先规定好的评分规则。

第五，它有一个决定考生下一道考题的选题机制。比内智力测验基于考生先前作答的表现来决定下一道考题。如果一个考生回答正确了某个年龄组中的大部分考题，那么后面的考题就有可能来自一个更高的年龄组；但是如果回答错了大部分考题，那么后面的考题就可能来自一个略低的年龄组。

第六，它有一个终止规则。比内智力测验在确定考生的最高水平(Ceiling Level)与最低水平(Basal Level)之后，就会停止测验。最高水平指的是考生全部回答错误的那个年龄组；最低水平指的是考生能够全部回答正确的那个年龄组。这个考生的真实水平就在这两个年龄组之间。

第七，考生的最终成绩由回答正确的考题决定。具体计算方法是 IQ 成绩是回答正确作答考题的加权 and，权重是年龄组。

图 1-1-1 是比内智力测验的图示。考题按照智龄(Mental Age)分组，每个年龄组内的考题由那些本组考生有 50%可能回答正确的项目组成。

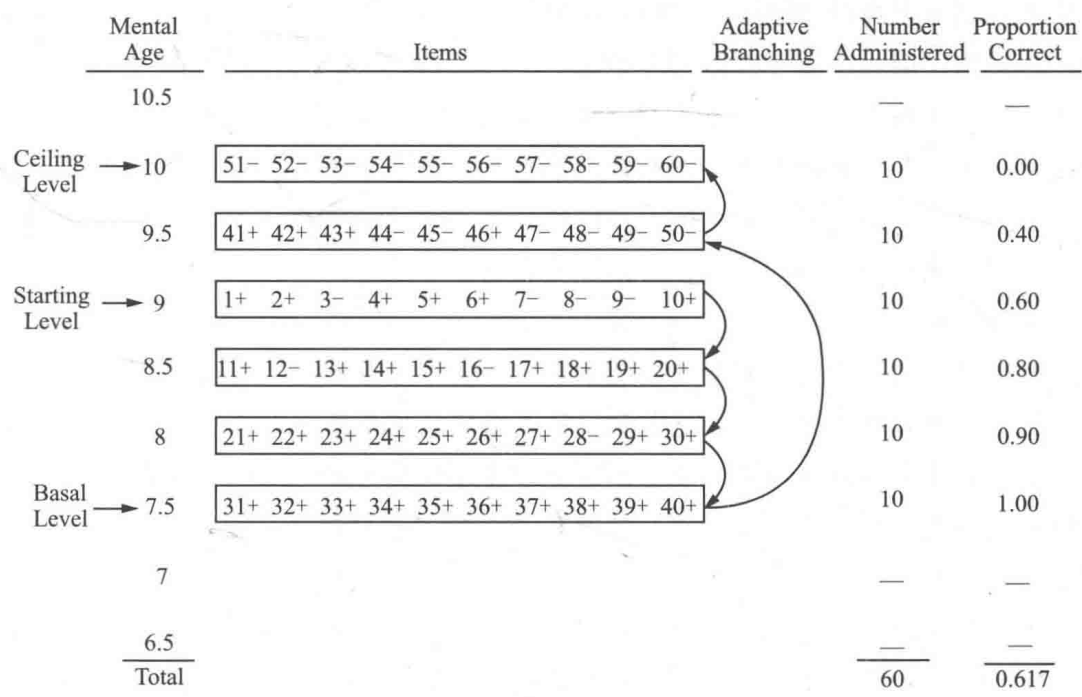


图 1-1-1 比内智力测验施测过程示意图

在这个图示中，考生从 9 岁年龄组的考题开始，他正确回答了第 1，2，4，5，6 和 10 题，错误地回答了第 3，7，8 和 9 题。因为没有全部正确或者错误回

答这些问题，因此9岁并不是这个考生的最高组或者最低组，考试需要继续进行。

此时，考生可以接受更高或者更低一组的考题。考官决定先寻找最低组，因此考生开始回答8岁半组的考题。考生正确回答了80%的考题。于是考官施测8岁组的考题，考生正确回答了90%的考题；接着施测7岁半的考题，考生全部回答正确。因此，这名考生的最低组被确定为7岁半。

用同样的方式，考官继续寻找这名考生的最高水平。考官首先施测了9岁半组的考题，考生回答正确了40%的考题；接着施测10岁组的考题，考生全都没有回答正确。因此，10岁组是这名考生的最高组。

二、自适应测验的主要特征

比内智力测验实测的例子展现了自适应测验的几个主要特征。

第一，每位考生有不同难度的初始题。理论上来说，比内智力测验可以根据考官收集的考生信息，从任何一个年龄组的考题开始。在例子中，如果考官从7岁半到10岁的任何一个年龄组开始，考生会接受同样的考题，得到相同的考试结果。如果从这个范围之外的年龄组开始，只是会增加一些考题，延长考试，但是考试结果不受影响。例如，如果考试从7岁组开始，考生应该会答对所有的问题，就会多找出一个最低水平组。同样，如果从10岁半组开始，考生就会多找出一个最高水平组，因为这个年龄组的考题比10岁组的考题都要困难。

第二，在收集到足够的考生能力水平信息之后，考试就会终止。在比内智力测验中，如果考题不能再提供任何新的信息时，测验就会终止。比最低水平组更简单的考题，对考生来说太简单了，而比最高水平组更难的考题又太困难了。这些考题都不能提供更多的信息，因此施测它们没有任何的意义。

第三，每个考生的考题数量可能会不同。在一个设计良好的自适应测验中，一般都会规定好考生能力测量的精度水平。在收集到足够的信息之前是不会终止测验的。在比内智力测验中，这个测量精度由最高与最低水平组确定。

第四，每个自适应测验可能会使用题库中不同的考题。自适应测验的突出特点就是从预先标定好的题库中选出最符合考生能力水平的项目进行施测。在这个例子中，这个考生回答了7岁半组到10岁组的考题。另一个考生很有可能回答5岁组到7岁半组的考题，而其他的人有可能回答8岁组到13岁组的考题。

第五，在自适应测验中，考生回答的理想考题的难度在50%左右，因为这个难题的题目能够提供最大的信息。在这个例子里，考生正确回答的考题比例是60%。这种“自我调节”的选题机制会使各类考生获得比较类似的心理体验。低能力水平的考生会觉得自适应测验比传统的纸笔测验简单，因为在传统的纸笔测验中他们会遇到更多的难题。相反，高能力的考生会觉得自适应测验比传统的纸笔

考试难，因为他们会遇到更多的难题。

第二节 计算机化测验

计算机化测验指的是以计算机为平台，向考生呈现考题的形式。计算机化测验有广义与狭义之分：广义的计算机化测验叫作基于计算机的测验(Computer-Based Testing, CBT)，狭义的计算机化测验叫作计算机化自适应测验(CAT)。CBT 包括所有以计算机为呈现平台的测验，例如，把纸笔测验的内容直接转化为计算机呈现，那么它就是一种计算机化测验。但是显然这样简单照搬纸笔测验内容的计算机化测验只是单纯的考试平台变化，不具备自适应的智能化特点。与此不同的是，CAT 不仅有测验平台的变化，还包括在第一部分描述的自适应特点(不同初始题、自动终止等)，因此成为目前最受欢迎的测验形式之一，也是本书的研究对象。在这一部分中，我们将简要地描述 CBT 的发展与特点，为以后各章提供一些背景信息。

信息技术的蓬勃发展给现代社会带来了深刻的影响。心理测量领域也被打上了现代信息技术的烙印，正在经历一场技术革命。心理测量的任务是在有关的量尺上给出测量对象的准确位置。这个任务非常复杂，需要耗费大量的人力与物力。现代电脑信息技术可以大幅提高测量中的工作效率，甚至使原来不可想象的任务变得轻松简单。一个简单的例子就是使心理测量可以变得快捷。由于计算机计算能力的大幅提升，它可以轻松地实施测验，包括呈现考题、收集作答、自动评分等，考生在信息化时代已经可以马上得到考试成绩甚至有关的诊断报告。

计算机信息化的巨大变化至少可以体现在两个方面。第一，计算能力的巨大提高。现在，一台个人计算机或平板电脑、智能手机等小型手持设备的计算能力已经远远超过 20 世纪的大型主机。著名的 ENIAC(Electronic Numerical Integrator And Calcula)可以占据一个很大的房间，可是它的计算能力却比不上一台普通的台式个人电脑。而更令人诧异的是，由于云计算的发展，在网络的帮助下，普通的智能手机或平板电脑也具有实施教育测量的能力。第二，各种个人计算机与简易手持设备的大量普及。计算机发展的早年，它以大型主机(Mainframe)的形式存在，是只有大型研究机构和企业才能负担的“奢侈品”。但是，目前个人计算机已经非常普及，甚至已经慢慢被更加便捷的智能手机与平板电脑取代，成为非常普通的日常消费品。

信息技术的革命性变革也不可避免地带来了心理教育测量的变革。早在个人计算机出现的初期，研究者就开始注重这些技术在测量中的应用。计算机化测验的研究始于 20 世纪 70 年代。自 20 世纪 80 年代以来，陆续诞生了一批大型计算