LANGUAGE TESTING

语言测试中的计量学原理

席仲恩 /著

SOME METROLOGICAL CONSIDERATIONS

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

语言测试中的计量学原理

席仲恩 /著

Language Testing:
Some Metrological Considerations

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

图书在版编目(CIP)数据

语言测试中的计量学原理 / 席仲恩著. -- 北京 : 社会科学文献出版社, 2018.8

ISBN 978-7-5201-2823-0

Ⅰ.①语… Ⅱ.①席… Ⅲ.①语言-测试-计量学-研究 Ⅳ.①H09

中国版本图书馆 CIP 数据核字（2018）第 109763 号

语言测试中的计量学原理

著　　者 / 席仲恩

出 版 人 / 谢寿光
项目统筹 / 祝得彬
责任编辑 / 仇　扬　郭锡超　高欢欢

出　　版 / 社会科学文献出版社 · 当代世界出版分社（010）59367004
地址：北京市北三环中路甲 29 号院华龙大厦　邮编：100029
网址：www. ssap. com. cn
发　　行 / 市场营销中心（010）59367081　59367018
印　　装 / 三河市东方印刷有限公司

规　　格 / 开 本：880mm × 1230mm　1/32
印 张：12.75　字 数：284 千字
版　　次 / 2018 年 8 月第 1 版　2018 年 8 月第 1 次印刷
书　　号 / ISBN 978-7-5201-2823-0
定　　价 / 78.00 元

重庆市教育科学“十二五”规划2011年度
教育考试研究专项课题项目

重庆邮电大学科研基金项目

前　言

本书主要是面向四类读者而写的——大规模教育考试试卷的开发者、语言测试研究者、语言教学定量研究者、语言测量结果的使用者。本书的目的是，用测量学的话语体系来分析语言测试或标准化大规模语言考试的过程。具体而言就是，把传统的语言测试或标准化语言考试过程分成三个阶段来看，第一阶段是测量工具的研发，第二阶段是测量工具的使用和测量结果的取得，第三阶段是测量结果的解读和使用。本书讨论的核心是三个阶段都涉及的量化问题以及量的测得结果的表述问题。

本书是我研究语言测试问题二十年的心得，是《语言测试分数的导出、报道和解释》（2006 年）一书的继续。那本书中，我只是尝试从计量学的角度看语言测试问题，而本书中，我把语言测试中的问题系统地翻译成计量学语言。那本书的核心是“难度”和“分数”，本书的核心是“量”、“测量”和“测量不确定度”。本书中，我努力传递三个重要信息：（1）量化问题是经验性问题，需要用经验或实验证据来回答；（2）包括语言测量在内的教育和心理测量包含了很大的不确定性，在结果的表

述和解读中必须充分考虑；（3）教育测量和心理测验需要用一套定义明确的术语和词汇来重新讲述自己的故事。

在本书的撰写过程中，我无时不感觉到自己汉语表述能力的不足。为了确切和明白，我尽量使用意义明确的简单词汇，尽量避免使用修辞手段，特别是拟人和比喻。例如，在本书中，我没有用“笔者”来指代我自己，也尽量不用“国家”一词，而改用其他意义更加明确的词语或表述方式。在学术著述中，“我”和“我们”也许是最难使用的两个词语。在本书中，“我”指的是“本书作者席仲恩”，“我们”指的是“本部分的作者席仲恩和本部分的读者”。在需要承担责任之时，我选择了使用“我”作主语；在邀请读者一起参与或分享之时，我选择了使用“我们”作主语。

本书是“重庆市教育科学‘十二五’规划2011年度教育考试研究专项课题项目”〔2011-KS-027〕和“重庆邮电大学科研基金项目”〔K-29〕的部分成果。重庆邮电大学外语学院的雷雪梅老师承担了这两个项目的秘书工作，从而使我专心致力于本书的撰写。社会科学文献出版社的仇扬等编辑对于本书的文字提了很多建设性意见。对于他们的贡献，我非常感谢。

限于我自己的知识和视角，本书中难免有不少缺点和错误，敬请各位读者批评指正。欢迎任何形式的批评，可以把批评意见直接发到我的邮箱（zhongenxi@126.com），也可以通过其他公开的方式。

席仲恩

二〇一八年春于重庆邮电大学紫竹园

目　录

第一章

绪　论

测试、测量、评估、评价，这是语言测试中经常被混淆的四个概念。作为全书的开篇，我们就先讨论这四个基本概念，然后再在此基础上讨论语言测量与语言有关决策之间的关系，最后讨论语言测量与语言研究之间的关系。

第一节　测试、测量、评估、评价

测试、测量、评估、评价，这是四个既有联系又有区别的基本概念，是必须分清楚的概念。

一　测试

本小节先尝试着定义“测试”，然后再讨论语言测试的名称与实质。

1. 测试的定义

先举几个例子。在注射某些药物（如青霉素）之前，先要给病人（通常是手臂内侧）皮下注射一定量的药物，看病人是否对拟用药物有敏感反应。过一定时间后，如果注射部位有皮疹，就做出接受测试的对象（受测）对该药物过敏的推断，因此不能使用试验药物。如果注射部位没有皮疹反应，就做出受测对该药物不过敏的推断，因此可以使用试验药物。要检验一个人是否尿糖过高，可能患有糖尿病，可以把特制试纸的一端浸入接受测试的对象（受测）的尿液里一段时间，然后根据试纸的颜色变化情况来判断受试尿糖浓度的高低。同样，要检验特定水源水的酸碱度（pH 值），也可以取少量的水样，把特制试纸的一端浸入水中一定时间，然后根据试纸颜色的变化来断定该水源水的酸碱度。一个供水系统或供气系统安装好之后，要进行加压测试。同样，高速铁路在投入正式运营之前，也要进行测试，而且测试的速度通常比正式运营时的速度要快很多。

分析一下这些例子就不难发现，所有的测试都涉及一个或几个标准，一旦达到或超过这个标准，就可以得出相应的结论。第二个发现是，同一种测试的条件都是标准化的，要么用同样的药剂、相同的时间，要么用同样的试纸，要么施加相同的压力，等等。第三个发现是，测试时的条件可能和常态时的条件相同，也可能不同。例如，测水的酸碱度时，就取正常的水；药物过敏试验的用药量就远小于正常用药时的量；测试供水供气系统时，气压或水压要明显高于正常供气供水时的压力。第四个发现是，下结论所根据的标准可能是定量的（如水压、气压、火车速度），也可能是定性的（如变色，出现皮疹）。

不难看出，在以上的四个发现中，第一个发现是最基本的，是测试的定义性特征。

作为测试的一个种类，语言测试也需要根据特定的标准，而且该标准既可以是定性的，也可以是定量的。到底使用何种标准，这要由决策的内容或性质决定。如果是判断受测是否能听懂或读懂一点某种语言的材料，定性的标准就可以了；如果要判断受测在多大程度上能听懂或读懂某种语言的材料，一般需要定量标准。尽管如此，在语言测试中，通常用的都是定量标准。而且，由于语言测试的对象通常是具有一定心理的人，所用的刺激材料也几乎不是自然的语言材料，测试的环境也很难是自然的真实语言使用情景，所以测试结果的不确定性难免会很高。

根据以上的讨论，我们可以试着给测试先下个定义：测试就是给接受测试的对象（受测）在规定的条件下施加一定的刺激，然后根据受测对于特定刺激所做出的反应，参照事先设定好的标准，做出受测是否达到这个标准的结论。如果受测达到了标准，就说该受测具有某种（些）属性；如果达不到标准，就说受测不具有某种（些）属性。

2. 语言测试的名与实

语言测试是英语 language testing 的翻译。这里的语言通常指第二语言或母语之外的语言，也包括外语；这里的测试通常指一个学科门类，也可以指一项活动、一组行动或一个过程，但不指一个项目。在语言学领域，语言测试是应用语言学的一个分支。同时，语言测试还是心理测验或教育测量的应用部门（参见 Chalhoub-Deville & Deville，2006）。

心理测验的英语名称是 psychological testing，汉语中也有人用心理测量或心理测量学（如郑日昌、蔡永红、周益群，1999；漆书青、戴海崎、丁树良，1998）。教育测量（学）的英语名称是 educational measurement，也有用 educational testing 的，例如，1985 年版、1999 年版和 2014 年版的美国《教育和心理测验标准》（*Standards for Educational and Psychological Testing*）都用了 testing 一词，之前的标准用的是 test 一词。这里的 measurement 用词并不准确，实际上是 testing（测试）之义。只不过，在汉语的心理学语境中，一般用“测验”，而不用“测试”。测验的意思是，根据测量结果进行检验。

严格地说，“测验”比“测试”更准确。为什么呢？这就涉及汉语中的检验和英语中的 test 一词。

先说汉语中的“检验”。这里的检验指的是统计检验，最常用的是其中的 *t* 检验（*t*-test）和 *z* 检验（*z*-test）。在包括语言测试在内的心理测验或教育测试中，根据分数对受测做出是否达标的推断，或者两个受测的分数是否显著不同的推断，或者受测甲的分数是否显著高于受测乙的分数的推断，这些都是明确的统计检验。

从统计检验的角度看，根据分数进行决策不仅必要，而且非常重要。其原因是，这样不仅有助于我们理解测试的本质，也有助于我们发现分数使用中的错误和问题。例如，在根据定量标准进行推断时，应该用 *t* 检验，但实践中通常用了 *z* 检验。至于为什么要用 *t* 检验而不应该用 *z* 检验，我们在第四章的“分数解释”部分再讨论。

现在再谈谈 test 一词。在英语中，test 不仅有统计检验中的

检验之义，还有两个用法与我们的讨论有关。一个用法是指考试或测试项目，另一个用法是指一组考试或测试用的刺激，即构成一套试卷的所有题目和说明文字的总称，如美国的托福考试、中国的大学英语（关于大学英语四、六级第一代考试的试卷构成和题型，参见杨惠中，Weir，1998）和英语专业考试（关于英语专业四、八级第一代考试的试卷构成和题型，参见邹申，1998）都是测试项目。前者的英语全称是 Test of English as a Foreign Language［英语作为外语的考试］，后两者的英语全称分别是 College English Test［大学英语考试］和 Test for English Majors［英语专业考试］。作为试卷意义上的 test 一词，在测试学中还有两个同义词：scale［量表］和 inventory［清单］。

需要指出的是，无论是 test、scale，还是 inventory，严格说只是一组刺激，而不是测量工具或测量系统的全部，有时甚至都不是测量工具或系统的核心或关键。例如论述型题目、自由作文或语篇翻译，其中的评分标准、评分人或评分软件，在测量工具或测量系统中，都明显具有比题目更核心、更关键的作用。关于测量，本章第二节有专门讨论。在结束本小节前，我们需要对语言测试下一个更加完善的定义。

语言测试就是给受测在规定的条件下施加一定的刺激，然后根据受测对这些刺激所做出的反应，参照事先设定好的标准，做出受测是否达到这个标准的结论。如果受测达到了标准，就说该受测具有某种（些）语言属性；如果达不到标准，就说受测不具有某种（些）语言属性。或者对不同受测个体或团体进行比较。如果所定的标准是量，或者所做的是量的比较，那么，就要对受测对于刺激的反应结果加以量化，再使用量化结果以

及结果的不确定性信息，对受测做出是否达到标准，或者是否有显著差异或显著强弱的统计推断。

可见，如果所用的标准是量，或者所做的是量的比较，那么，测试就包括了测量。作为一门学科，语言测试不仅包括语言测量工具的开发和使用，还包括根据测量结果进行的统计决策以及关于决策后果的研究。但要进行统计决策，或者要控制决策的错误率，那不仅需要测量结果，还需要与测量结果相伴随的不确定度，即语言测试传统上所谓的误差或标准误。如果用 t 检验的语言讲就是，只知道分子上的信息而不知道分母上的信息，t 值是求不出来的。

二　测量

汉语中的“测量”和英语中的 measurement 并不完全对应。前者只指一种过程，而后者不仅可以指过程，也可以指学科和结果。作为术语，在讨论计量问题时，测量最好只用过程这个意思。但不幸的是，在包括语言测试在内的心理和教育测量学的英语文献中，measurement 在这三个意思上都用，经常使读者混淆。

1. 测量的两个不同定义

根据国际标准化组织指南 99：2007《国际计量学词汇——基础通用的概念和相关术语》［ISO/IEO Guide 99：2007 *International Vocabulary of Metrology-Basic and General Concepts and Associated Terms*（VIM）］2.1 的定义，

> 测量是一个通过试验获得一个或多个量值的过程，该

> 所获量值很可能就是一个量的值（Process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity）。

但是，在社会测量文献中，一般都采用斯蒂文斯（Stevens，1946，p. 677）对测量的定义，即

> 从最广泛的意义上讲，测量就是根据规则给物体或事件赋数（Measurement，in the broadest sense，is defined as assignment of numerals to objects or events according to rules）。

这个原本引自英国物理学家兼科学哲学家坎贝尔（Norman Robert Campbell）的定义，把社会测量远远隔离在计量学之外，而且，使得社会测量至今也未能满足基本测量的必要条件（Resse，2017）。

2. 两个不同定义的解读与比较

ISO/IEO Guide 99：2007 的定义看似简单，实际上并不简单。在 VIM 2.1 中，测量是在定义了量值（也叫量的值或量）之后才定义的，而且后面还附了三条注解。

第 1 条注解就指出，测量不适用称名属性（Measurement does not apply to nominal properties）。这一条，就已经把社会测量（包括心理和教育测量）区分出来。因为，根据斯蒂文斯（实为坎贝尔）的定义，测量是适用于称名属性的。对比坎贝尔对测量的定义和 VIM 2.1 的定义不难发现两者之间的两个差别：VIM 2.1 定义所测量的是“量”，坎贝尔定义所测量的是“事物”或

“事件”；VIM 定义测量的结果是“量”，坎贝尔定义测量的结果是“数”。“量”包含了数，是数和单位或参照系的组合。但单纯的“数”并不能构成量。量化不是数化，量化的核心和基础不是数，而是单位或参照系。

VIM 2.1 对测量定义的第 2 条注解的意思：测量隐含了对多个量的比较，测量包括了对实物的计数（Measurement implies comparison of quantities and includes counting of entities）。换句话说，数一类实物的个数或同一个实物出现的次数也是测量的一种形式。不难看出，这条注解也有明显的局限。既然数实物是测量，那么数现象或者抽象的概念算不算测量呢？当然也应该算。打了几次雷，出现几次闪电，某个念头在脑海里闪现过几次，等等，这些都是测量的例子。

VIM 2.1 对测量定义的第 3 条注解的意思：测量预设了对所测的量的描述，该描述与对测量结果所拟定的使用相匹配；测量还预设了一套测量规程和一个校准过的测量系统（即测量工具），这个系统要按照特定的测量规程操作，包括特定的测量条件（Measurement presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions）。

这条注解的内容非常丰富。第一层意思是，它要求测量结果与结果的使用相匹配。换句话说，目的或用途决定了特定测量的恰当性或合适性。例如，如果结果的使用是一个人分几个苹果，那么，清点用来分配苹果的个数和参加分配的人数就是