席仲恩 著

语言测试分数的导出、报道和解释

Reliability is the degree to which the result of measurement for a given population of test takers attenuates the correlation between two variables. It is a population parameter measured by reliability coefficient. The geometrical mean of the reliability coefficients with respect to the measures of the two variables indicates the extent of attenuation.

The Derivation, Reporting and Interpretation of Language Test Scores

$$X_{ji} = U_{ji} \times H_{0i} = \frac{U_{ji}}{P_{0i}} C_0$$

$$\sigma(\hat{X}_{ji}) = \sqrt{H_{0i}^2 u^2(\hat{D}_{ji})} = H_{0i} u(\hat{D}_{ji})$$

$$X_{ji} = D_{ji} \times H_{0i} = \frac{D_{ji}}{\overline{D}_{0i}} C_0$$

$$\rho = 1 - \frac{\overline{u}_c^2(\hat{X}_{ji})}{\sigma^2(\hat{X}_{ji})}$$

$$H_{0i} = \frac{C_0}{\overline{D}_{0i}} = \frac{1}{\overline{D}_{0i}} C_0$$

四川大学出版社

席仲恩 著

语言测试分数的导出、报道和解释

The Derivation, Reporting and Interpretation of Language Test Scores

四川大学出版社

责任编辑:黄新路
责任校对:吴　昀
封面设计:米茄设计工作室
责任印制:曹　琳

图书在版编目(CIP)数据

语言测试分数的导出、报道和解释 / 席仲恩著. —成都: 四川大学出版社, 2006.9
ISBN 7-5614-3543-6

Ⅰ.语...　Ⅱ.席...　Ⅲ.语言-测试-英文
Ⅳ.H09

中国版本图书馆 CIP 数据核字 (2006) 第 116081 号

书名　**语言测试分数的导出、报道和解释**
The Derivation, Reporting and Interpretation of Language Test Scores

著　　者　席仲恩
出　　版　四川大学出版社
地　　址　成都市一环路南一段 24 号 (610065)
发　　行　四川大学出版社
印　　刷　郫县犀浦印刷厂
成品尺寸　140 mm×202 mm
印　　张　12.5
字　　数　305 千字
版　　次　2006 年 8 月第 1 版
印　　次　2006 年 8 月第 1 次印刷
定　　价　28.00 元

◆读者邮购本书,请与本社发行科联系。电 话:85408408/85401670/85408023　邮政编码:610065
◆本社图书如有印装质量问题,请寄回出版社调换。
◆网址:www.scupress.com.cn

前　言

中国是考试的实践大国，却是测试的发展中国家。考试不同于测试。前者通常是权力的施展、权威的展现；后者不过是工具的拓展，是公共服务的一个方面。测量是关于量的值的获取和获取工具制作的学问，检验是关于量的值的解释的学问，测试则是测量和检验的合璧。测试理论不仅可以指导测试实践，也可以用来指导考试实践，还能对考试实践进行校验；它不仅可以直接地服务社会，也可以通过服务于科学研究而间接地为社会效力。

本书是笔者对英语专业四、八级考试所提建议的理论基础篇，也是笔者博士论文前5章的拓展和改写。当时提建议的宗旨是，让四、八级考试这一强大的权力工具能够受到社会的有效监控，不至于被误用或滥用，从而为社会提供更全面、更优质的服务。考虑到论文的后3章涉及的问题过于具体，对于广大的考试研究和实践工作者、测试研究和实践工作者以及定量研究工作者的参考价值不大，成书时全部删掉。

拓展的目的无需赘言。现将改写的原因做个简要的交代。首先需要说明的是，此次修改，已经是第二次修改，而且改动很大，其中部分章节几乎重写。改动的原因是笔者在信度理论研究方面又取得了新的进展。原稿中信度部分用的基本是概化理论的思想，但笔者最新研究结果从多方面表明，概化理论可能存在重大的理论错误和应用误导。新的信度理论直接从信度的定义入手，根据现行自然科学计量理论，明确了定义公式中各个参数的意义，避免了量值不可获取的参数（即真分数方差和误差方差）。新理论应用不确定度传导定律，在个体水平上解决了经典理论一直无法解决而概化理论通过复杂的方差分量分解技术才能解决的不确定度分量合成这一难题。因此，在功能上替代了经典理论和概化理论。

新理论既不需要经典理论中的严格平行条件，也不需要概化理论中的内部一致性条件。新理论明确了个体测量结果的不确定度和团体总/平均不确定度之间的关系，解决了经典理论和概化理论无法解决的个体特定不确定度问题。此外，新理论涉及的数学知识也非常简单，只用到样本方差的计算公式，不像概化理论，既涉及复杂的研究设计理论，又涉及复杂的方差分量分解技术。

自从“信度”概念提出以来，学界关于信度到底是什么的争论一直没有休止。为了与自然科学界的测量学接轨，笔者建议：

用不确定度和变异系数表示测量或估计结果的信度；为了使所言和所指一致，用衰减指数替代传统的信度系数。原因是：传统的信度系数并不是一个能够单独反映测量结果信度的参数，也不是一个能够单独反映测量工具信度的参数，它所能单独反映的，正是根据这样的测量结果进行相关分析时，相关系数受到衰减的程度。例如，有一组数据，据此计算出的所谓信度系数是 0.85，那么，这个系数的确切含义是：用信度系数同为 0.85 的两组数据做相关分析，将会把本来的相关系数衰减到它的 85%。如果两组数据的所谓的信度系数值不等，则它们的几何平均数就是衰减系数。那么，学界为什么一直把衰减指数的这个概念叫做信度系数呢？因为用语不慎，斯皮尔曼（1910）把这个概念称作信度系数，而他真正所指的却是衰减指数。

此次大修改正是为了明确信度系数的含义，限定信度系数的应用范围。为此，笔者不得不把传统的笼统的测量过程分解为测量和抽样检验两个具体阶段。信度系数仅对测量阶段有意义，在抽样阶段，无需定义信度系数。书中还记述了其他方面的探索结果以及探索过程中笔者的一些思考。不过，所有这些争论和思考，已经成为过去。如果书中的某个公式被确认，或者某个看法被证实，那定是上苍的恩赐、先贤的施舍。书中的看法，均系成书时

笔者的看法，自然也是成熟的看法。但是，成熟不等于正确，成熟仅意味着明确，而且仅仅是对于提出者而言的明确。对于学界而言，所提看法是否明确，是否正确，还有待各方同仁的评判。

在本研究的各个阶段，我得到了多方面的慷慨援助和热情帮助。我的导师邹申教授，给我布置了这项重大的研究任务。不然，我可能不会系统地研究分数问题，自然也不会有这本著作。美国“自然科学基金会计算机及网络系统分会”的赵伟主任，不仅为本研究免费提供了大量的最新资料和历史资料，还多次和我讨论本研究的有关问题。周越美博士、余国兴博士、唐雄英博士，都利用各自在国外的宝贵机会，向本研究提供了不可或缺的资料援助。美国衣阿华大学“测量与评估高级研究中心”的 Brennan 主任曾回答过本研究中遇到的一些问题。芝加哥大学“客观测量研究院”的 Linacre 教授不仅及时回答笔者所咨询的问题，而且总是在第一时间把“Rasch 通讯”的信息通知我。绍兴文理学院外语学院在房间非常紧张的情况下，给我安排了专用办公室；英语系也在教学任务非常繁重的情况下，减轻了我的教学任务，使我有更多的时间投入研究工作。浙江省社会科学基金和绍兴文理学院对本研究提供了资助，使得本研究得以顺利进行。绍兴文理学院学术专著出版基金对本书的出版进行了资助，四川大学出版社的

黄新路先生为本书的设计也花了很多心血。对于这些援助、帮助和资助，我谨在此表示衷心的感谢。

我的妻子赵泓女士，不仅一直支持本研究，也直接参与了本研究的部分工作；我的女儿席温纳也为本研究提出过不少批评意见。由于研究工作，我长期未能对妻子、女儿尽到家庭责任，对于她们的支持、宽容、理解和谅解，我表示最崇高的敬意。

席仲恩

2006年5月

Preface

This monograph deals with the derivation, reporting and interpretation of test results in the broad sense. It is the revision and extension of the first five chapters of my doctoral dissertation, *The Derivation, Reporting and Interpretation of Language Test Scores with Some Recommendations for the TEM*. The original work consists of eight chapters. It forms part of an attempt to unify the theories in educational and psychological testing in general, and language testing in particular, with those in the hard sciences. The motivation of the attempt is to simplify the current theories so as to prevent their misuses and abuses, and ultimately facilitate their applications.

In preparing the dissertation, my goal was specific: to offer some practical recommendations for the TEM testing program with respect to the expression, communication and consumption of the TEM results. I came up with some specific recommendations first, and then started searching for theoretical supports for each of them. But unconsciously, as I realized later, I slipped into validating the complicated theories needed for supporting the recommendations. I was shocked again and again as I tried to disentangle the current theories, for I found that many of the conventional theories were unjustified. To compensate for the urgent need, I developed some models of my own, and suggested that these models along with the new item hardness model I built a few years ago be used. All these models together with the recommendations for the TEM testing program were recorded in the dissertation.

As I felt these models may be of some general significance, I

decided to take out the first five chapters and have them so revised that they may be self-contained and independent enough for publication. By early November, 2005, the revision work was completed. So I laid the manuscript aside and began to look for a publisher. In the meantime, I reflected over the theories that I have invented and the theories that I have advocated. Thanks to the recent debate on reliability theory waged by *Educational and Psychological Measurement*, my special attention was drawn to reliability issues. The more I thought over the reliability theory I have used in the book, the more uneasy I became. In the late fifties of the 20th century, Cronbach and his associates thought reliability was the most mature part of testing theory. But as they started to work with it, they realized that reliability theory was far from being mature and error free. To liberalize and extend the classical theory, they developed generalizability theory (GT). This theory assumes, among other things, that the input data be linear or at the interval or ratio level. Unfortunately, all the data used by GT theorists are clearly nonlinear. As the item hardness model I have developed is meant to transform the nonlinear data into linear data on the performance dimension, I translated the raw data by weighting each item with its own hardness, and then tried the data on the simplest GT design $p \times i$. What a surprise! The result was exactly the opposite of what GT predicts: *the index of dependability was greater than the coefficient of generalizability.* I tried several real data sets, the results were exactly the same. To pin down the problem, I looked at the next simplest design $i:p$. I found that the interpretation in GT contradicted the conventional sampling principles on the one hand and was logically incoherent on the other. A real shock to me! According to sampling theory, $p \times i$ design and $i:p$

design should be equivalent in effect, in terms of both population estimate and the uncertainty of estimation, but GT claims that the overall group uncertainty is less for the former design than it is for the latter. When the uncertainty of group mean is in question, GT claims just the opposite! How could this be possible? In interpreting individual measures, *A* is greater than *B*, while in interpreting group mean measure, *B* is greater than *A*? As group mean is based on individual measures, *A* must be equal to *B* for interpreting both the individual measure and the group measure. Therefore, there should be no distinction between the *p*×*i* design and *i*:*p* design, at least when the study of uncertainty is the concern.

As Cronbach and his associates referred us again and again to R. A. Fisher (1925) for decomposition of variance components, I tried to find that old book to see if there is any solution in the classic. Thank God! In chapter 7 I bumped into Fisher's basic definition of intra-class correlation and learned why he applied ANOVA to intra-class correlation analysis. From the perspective of the basic definition, another contradiction was identified with GT: According to the well known relationships in probability theory with respect to variance and covariance, it can be easily shown that like inter-class correlation, intra-class correlation, too, is invariant of the linear transformation of input data, that is, *the dependability index should be identical to the generalizability coefficient.* But according to GT, the linear transformation of the input data will turn the dependability index into the generalizability coefficient, with the latter being equal to or greater than the former in value, and one being different from the other in function. Although GT claims that generalizability is for relative decision making, whereas dependability is for absolute decision

making, and yet what it really means is that dependability is calculated from the non-transformed measure (raw score) and generalizability is calculated from the transformed score (deviated score or standardized linear score). The contradiction identified suggests that either GT is syntactically incorrect or the established mathematical relations are invalid. We tend to believe that the former is true. This being the case, the semantic distinction is unnecessary.

In the mean time, efforts had been made to revise the conventional reliability theory in strict accordance with the current GUM and VIM spirit: To avoid using the ideal quantities and focus the attention on the obtainable ones. To be more specific, instead of using the concepts of the true scores and error scores, we should focus our attention to the estimated scores and their associated uncertainty. The fruit of this endeavor is the basic definition form of reliability coefficient. I was quite happy with the result, for with the basic definition, reliability coefficient can be calculated from the measure specific uncertainty resulting from both Lords' binomial sampling model and the information function of item response theory. If reliability is defined as the ratio of "true" variance to observed variance, the basic definition is the unconditional variant and operational definition of reliability. Besides, it makes explicit that only when the group variance is greater than or equal to the overall group uncertainty expressed as a variance, is the reliability coefficient defined. Seen from this perspective, intra-class correlation stepped up by Spearman's general formula is only one conditional approach to reliability. If the internal consistency condition is met and the test taker group is heterogeneous enough, that is, if none of the inter-item correlation is negative and the variance of the estimated measures of

the group is greater than or equal to the overall group uncertainty, GT offers a valid approach to reliability. If any one of the conditions is not met, neither GT nor classical test theory is valid for estimating reliability coefficient.

As is known to all, reliability coefficient alone cannot indicate the reliability of measurement result. Therefore, it is recommended, in practice, that both the reliability coefficient and group variance or its positive square root be reported. Not satisfied with the current practice in educational and psychological testing, I went on looking for better measures of reliability in GUM (*Guide to the Expression of Uncertainty in Measurement*) and VIM (*International Vocabulary of Basic and General Terms in Metrology*). There, the answer was ready for us. Both the uncertainty and the coefficient of variation (known as relative standard uncertainty) are good measures of the reliability of the result of measurement and they were exactly what I had been looking for. I was so convinced that I decided to recommend the GUM-proposed practice to educational and psychological testing, language testing included.

But what about the conventional reliability coefficient? Shall it be discarded or shall some specific functions be identified for it? I chose the second alternative. As reliability coefficient is conventionally used for three purposes, namely, estimating overall group uncertainty, estimating the individual specific true score, and correcting for the attenuation of correlation coefficient, I checked them one by one. With the introduction of the new theory, there is no point in using reliability coefficient to estimate the overall group uncertainty, for before the coefficient is calculated, the overall group uncertainty is already there. Then I checked it for the second proposed

use. In defining reliability coefficient as the square of the correlation coefficient between the true score and observed score, it is assumed that true score is independent of observed score. This is equivalent to saying that the error of measurement is a constant across the test taker population (see Gulliksen, 1950). But in estimating the true score, it is assumed that the farther away an observed score is from the group mean, the greater is the error. This contradicts the independence assumption. On the other hand, by definition, reliability coefficient characterizes only the random error, but in estimating the true score, it is used as if the error it characterizes were systematic, positive for scores that are lower than the group mean and negative for scores that are higher than the group mean. This is another contradiction. Evidently, the second use is invalid.

To test the validity of the third use, I traced back to the very origin of reliability coefficient, that is, Charles Spearman's original papers. In Spearman (1904a, 1904b, 1910), the original proposal was found: to correct for the attenuation of correlation coefficient caused by the faulty input data. But Spearman (1904b, 1910) made it clear that to correct for the attenuation, the "error" characterized by the coefficient must be due to observation rather than sampling. Spearman purposefully rejected the test-retest error not because its effect *should not* be corrected, but because the effect *cannot* be corrected. To him, repeated measures were measures of different functions rather than different observations of the same function (Spearman, 1910). In deriving the general formula, Spearman let, at a critical stage, the left side of the correction formula be equal to "1." This is the condition that must be met before reliability coefficient is used to correct for attenuation. It is easy to show that when sampling

over content is the case, the correction formula is incorrect. Apparently, according to the function the coefficient exercises, it should be called the *index of attenuation*. Unfortunately, owing to Spearman's original loose usage, it came to be called *reliability coefficient*, a name that fails to mean what it intends to mean. As *reliability* is a frequently used word, we take it for granted that it means what it usually means in our everyday language. But this time, *reliability* is not that reliable! As more and more evidence had been collected in the long search for the exact meaning of reliability, I felt obliged to rewrite the manuscript.

Into the following text, the foregoing discoveries and more have been integrated. But that was no easy job for me. On the one hand, the coherence and the practical orientation of the original text had to be kept. On the other hand, the new discoveries in theory had to be built in. In order to delimit the use of "reliability coefficient," the conventional process of testing has been divided into two phases: The measuring phase and the inspection-by-sampling phase. And reliability coefficient is only defined for the result of measurement, not for the result of sampling. Throughout the text, more attention was paid to the semantics than to the syntax of testing, for it is my conviction that testing itself is semantics rather than syntax, at least, it is more of the former than of the latter. Unlike the conventional approach, the approach taken in the text is individual based. That is, both the individual test taker's performance and the uncertainty of its estimate are taken as the immediate objects of investigation rather than the individual difference and the overall group uncertainty. It seems to me that individual difference is not qualified as a proper measurand, nor should the overall group uncertainty be regarded as a

as a parameter to be directly estimated. Instead, the former should be determined on the basis of the result of measurement or estimation, and the latter, if needed at all, should be evaluated through the individual specific uncertainty. Compared with the item response theory of latent trait, the theory in the following text may be referred to as the item response theory of apparent performance.

So many people have lent their helping hands at various stages of the present research that it is impossible for me to list them all, and their helps have come to me in so many different ways that I could mention only some of them. First and foremost, I owe my deepest debt to Prof. Shen Zou of Shanghai International Studies University, who literally assigned the task to me. Without her "assignment," I would not have taken such a systematic look at score related issues. I am also grateful to Dr. Wei Zhao, senior associate vice president for research of Texas A & M University and division director of the Division of Computer and Network Systems, National Science Foundation of the United States, who has not only timely bought and brought me most of the necessary reference materials at his own expense, but also has spent many precious hours discussing the central topics of the research with me. To the accomplishment of the present research, his generosity, help and especially his penetrating questions are indispensable. Dr. Yuemei Zhou, Dr. Guoxing Yu and Dr. Xiongying Tang, all have aided the present research by collecting important reference materials. Without their aids, it would have been impossible for me to know some of the key events in the history of reliability studies. Both Robert Brennan and John Linacre answered my questions relating to the history of reliability. John has also kept me being informed of the latest of the Rasch-model-based research

and other activities by sending me the *Rasch Measurement Transactions* the moment it is ready. I would like to express my heartfelt thanks to them all for their great help.

To my family, my daughter and my wife, I would like to say a special thank you for their sacrifice, criticism, concern and help. My dear daughter Winnie, has not only suffered from the long absence of Daddy, but also offered quite a lot pungent criticisms to the presentation of theory. In the past seven years, my wife, Hong Zhao, has not only suffered from the scarcity of family activities but also the worries of my health. For her love, understanding, sacrifice and unfailing support I dedicate this book to her.

This research was largely carried out in 1999-2006. The work was supported in part by the provincial Social Science Foundation of Zhejiang Province, and in part by Shaoxing University.

Zhong'en Xi
Shaoxing, Zhejiang
May, 2006