从写作测验信度研究看开　　测

——

冯瑞龙

(□□□□□□□□□□□□□□□□□□□□□,□□ □□519085;□□□□□□□□□□□,□□ □□361102)

摘　要:□□□□□□□□□□□□□□□——□□□□□□□□□□□□□,□□□□□□
□□□□□□□□□□□□□□□□□□。□□□,□□□□□□□□□□□□□□□□□□□。
□□□□、□□□□□□□□□□□□□□□□、□□□□□□□□□□□□□□□□□,□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□
□□□□□□□□□□□□□,□□□□□□□□□□□□□□。□□□□□,□□□□□□□
□□,□□□□20□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□,□□
□□□□□□□□□□□□。

关键词:□□□□□;□□;□□□□□□;□□□□□□;□□□□

中图分类号:G449　　**文献标志码**:A　　**文章编号**:2221-9056(2014)02-0160-08

□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□、□□□□□,□□□□□□□□□□□□□、□□□□□,□□□□□□□□□过□□帮助□□□□□帮助□□□步□□□。□□□□□□□□□点还□它□□□□□□多□□□□□□□武□□。□□□,□□□□□□□□□□□□□□□□又□□□□□劣势,除□□分/级□准难□厘□、过程费□耗□□□,□严□□□□恐怕就□□□□□□难□□□□□□想□水平□。□□□不仅□衡□□□质□□　□□□□□□□□,□□□□　□□□□一　□□　□　。□　□程□□,　□□□□　,□□　□□□□□□□不　严　、□,　□□□□□□□□□□□□□□　。□,　□□□□　□□□　□□□　□不□□步,□□,□□　□　□□□□

收稿日期:2014-02-17

作者简介:冯瑞龙,男,澳大利亚格里菲斯大学文学博士,北京师范大学香港浸会大学联合国际学院副教授、博士生导师,研究方向　　文学。
　,男　国　大学教　学博士,　大学　教　学院副教授　　学　博士生导师,研究方向　国际教　。Email:zhuyu@xmu.edu.cn

基金项目:UIC Research Grant(2012-13)R201307　教　学　国　研　。

□□□□□□□□□□□□□□□□□□□□□□□□□□□。

□□□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□——□□□□,□□□□□□□□、□□□20□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□。□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□、□□□□,□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□。

□□□□□□□□□□□□□,□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□——□□□□□(authentic testing)□□□□□□□□□。□□□,□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□?□□□□□□□□□□?□□□□□□□□□□□□□□□□□□□□□□□□□□□□□?□□□□□□□□□□□□□□□□□□□?□□□□□□□□□,□□□□□□□□□□□□□□□。

二、信度理论流派的简介与比较

(□)□□□□□□(CTT)□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□。□□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。

□□□□□□□□□(□ Inter－rater Reliability □ Intraclass Correlation Coefficient,□□□□□ICC)□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。ICC□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□,□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。□□,□□□ICC□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□,□□ICC□□□□□□□□□□□□(Wikipedia,2010)。ICC□□□□□□□□□□□□(□□□□□□□□□□□□)□□□。□□□□□□□□□□□□□□□ $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ □□□。□□ Y_{ij} □□□ i □□□□□□ j □□□□□□,μ □□□□□□□□,α_i □□ i □□□□□□□□□□□□□□,ε_{ij} □□ i □□□□□□ j □□□□□□□□。

□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□。□□□□□□□□□ICC□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□(Julius Sim & Chris Wright,2000:335)。

(□)□□□□□□(IRT)□□□□□□□□

Linacre (1989) □□□□□(George Rasch, 1980)□□□□□□□□□□□□□。□□□□□□□□□□□□□□□□□□□□□□□,□□□□□□□□□□□□□□□□□□□□□□□□□、□□□□□□□□□□□□□□□□□□□□。

□□□□□□□□□□□□□□□□□□□□□□□、□□□□。□□□□□□□1。□□、

外 合 0.5 1之 时,测评是富 的。当 大 1.5 2时, 测评 ,但 。大 2时,测评 是 信的。

型 用 性 能。这 性 能, 题 的 式 能 当的其 ; 性 能, 能 当的 提 , 一 的 对 同 其 的评(George Engelhard,2008); 性 能, 题 度 同的 提 , 题 的 其 外 题 的 (George Johanson & Abdalla Alsmadi, 1998)。是 存 性 能是 过 之 的 用的 的。 一 用 的 性 能 予一 偏 ,当 Z 的绝 大 2时, 一般认 的 性 能 。 存 性 能则暗 的测 题 能 待 ; 性 能 则提 能 了较大偏 ; 的 性 能则能 映 测的 能 掌 的 。

信度也是 型涉猎的范畴之一。 用 型的 据 能提供 信度统 量, 一 是 范围 0 1之 的 割 信度, 一 是 范围 1 穷之 的 割比。尽管 型 一 的 能 这 种 ,但 的这 大 , 其 的这 则 其 各 中各要素带 的 性 能 。倘若这 各要素 的 性较 则能 映测量 较高的 性。但 各要素(一)之 性较 , 映的 是评 属 一种 理 的情况。

( ) 理论(GT)的信度评

理论认 测 的方 是 方 量的。这其中既 希望过 测 的 的 造的 的方 量(称 测量 方),也 扰 素造 的方 量。采用 理论的 G 据 能 这 方 量 。测量 方 量 方 中 的比重 大 靠。 了提高 的 推 性 靠性, 要 过控制降低比重大的 扰性方 量, D 则能让 G 的 了 变 扰变量的 量, 变评 、题的 量方 会造 测 信度的何种变 , 帮 能保证理 信度的最佳 测 评 措 。

理论 信度时, 了 照 照测 的信度 。 g , Phi 。这 种 时 测 象 的方 子, 方 也是 母的 之一。 一 g 言是 方 , 测 象 G 的各 的 用造 的方 量之 。 Phi 言,个是绝 方 ,是 了测 象方 量 外的各 方 之 。

( )信度 的 评

经典测量理论 的信度 最大的 用经典测量理论 型的 提 较低,实测 据 。 一是 个 能 方 之外的 一 方 量,也 认 控制的一 方 素, 何 测量的信度提 测 方 的 。外,其 用的 本 大,是 能 本的 。

理论的 型 经典测 理论的信度 理论信度理 最大的 同 个的 是 本的。 其着 个 信度的 是 测 的 信度, 是给 一 的 一 要素的测量信度(割 信度 割比)。若要 信度, 信度 理 的 。比 , 评 的信度 较低, 性评 能 , 何合理 评 、 评 控评 过

□ 。然□ □ □ □ □ □ □ □ □ □ □ 不□ □ □ □ □ □ □ 分□ 素□ □ □ 如 □ □ □ □ □ 。

□ □ □ □ □ □ □ □ 分析□ □ 虽然不□ □ 样□ □ 立,□ □ □ 取 □ □ □ □ □ □ 素□ □ □ □ □ □ □ □ □ ,□ □ □ 够□ 认□ □ □ □ □ □ □ □ □ 分□ □ □ □ ,□ 够□ 过全 □ G□ □ 推□ □ □ □ 素、□ □ □ 素 面□ □ □ □ □ □ □ □ □ 分□ □ □ ,□ □ 够□ 过 D □ □ □ □ □

□□更有□于提□写作成绩的可□性和概化□□。

□有研究□同时使用了概化理论以外的□□□具。□:Sudweeks □□(2004)的□□研究使用概化理论和□□□□□型同时估算了写作评□□在的误差源和写作成绩的信度,并据此提出了改善评□□□的□□。结果□□写作题目以及被试与题目的□□作用比较□,而评卷□和考试□□□成的方差较低。这也□□着□加写作题目是改善此项写作测试成绩可□□性的有□□□。

Schoonen (2005)以 G 研究估算了被试写作□□、作文题目、评□的项目(内□或□□□用□)以及评□方式(□体性评□□或□□性评□)的□应,并通□结□方□□型估算了写作□□的方差成□。□研究的被试是 89 □ 6 年□学□,□□被要求写□□作文,□□作文的内□□□及□□□用□个方□被 5 □评卷□以□体和□□□种方式评□。□□结果□□写作成绩的可□□性以及评卷□和写作题目的□应在□大□度上□□于评□方式与评□项目。□体而□,写作题□□的方差要□于评卷□□□的方差。

在国内,□□□和□□□(1998)的研究□□□发现写作题目对测试成绩可□□性的□应,但发现了□同文体对评□误差有重要□□,□论文的评□误差最大。□□、□□□(2008)□用□□概化理论对出国□学□□□□□测试的 30 □□试的写作□□进□□□,测试有□个写作□□,□□ 1 要求□试发□□□力,□据提供的□□自□□写□□;□□ 2 是□试较□□的□题作文,□□□述自□的学□与□□经历、□□□好□。评□□用 Jacobs □□ 1981 年设计的□□作文评□量□“ESL Composition Profile”。研究结果□□□个写作□□的合成总□的评□信度较□。

可□,写作题型和题量□是□□写作成绩可□□性的要□。□题型而□,□些新题型□□写写作或□写写作成绩的可□性并□低于□□的□□写作,考□到这些题型更□合真实测评的□□,□□写作□□测试也可以□□□用此□题型为□题写作的□□。而□所测的□体来□,对□论文的评□可能相对更主□一些,其成绩的可□□性会相对□差。□合考□被试的□□□□□,或□在□、中□□□写作□□考试时,应□□强□考□写□论文体的作文。至于题量方□,所参考的文□□□一□□□相关测评应有□□或□□□以上的写作试题,以□□被试写作成绩的可□□性。

(□)关于评卷□

□□上述研究□或□或□地□□□加评卷□□□□是提□写作成绩概化□□的有□□□,□□之,评卷□一□□是写作成绩方差的主要来源,但写作评□□□中,□同评卷□对同一□作文评□的差□□是□□了□内一些学□的□□。例□:Johnson 及其同□(2005)□以概化理论研究了□□评卷□评□□出现差□应□□□理的问题,□理□问题的□同方法会对□作性评□的信□度□□□同□□。具体而□,Johnson 及其同□比较了以评卷□的□□□为被试最终□□和通□□论□□一□性评□□种□同方法所□被试写作成绩的□确度,并考□了□论□□中是□会出现个□评卷□□于□□性地□的□□。研究结果□□,以计算□□□或□□论解□评□差□的□果并□有□□差□,□□对于提□评□□确性□无□大□□。在进□□体性评□时,□以□论方式解□评□差□,相对更□□出现个□评卷□□于□□性地□的□□。

相较□方的研究,国内的相关研究更□地发现了评卷□对于作文成绩可□□性的□□□应。例□,□□□和□□□(1998)使用概化理论□□了 6 □评卷□对 20 □学□□□三种文体的作文进□□□性评□的□据。结果□□在作文评□中,评卷□□应最大,题目□应□□□。

此外,国内的一些相关研究□评卷□的评卷经验和□□学□□也考□在评卷□□应当中。□□□(2010)以个案□□的方法对新□评卷□在 HSK 写作测试的评卷□信度进□考□,研究从 2009 年 4 □的 HSK(□□)写作测试评□的 48 □评卷□中□□□□了□□(□□中一□一新□□评卷□),□用概化理论对□□的评□信度进□了□验,结果□□□□的测验信度较□,同时了解到新□评卷□在对评□标□的□□中存在的差□。□□(2005)□用概化理论对有、无□□学□□

的评卷　给 20　试□级□□,　试(HSK)　分之　文评分的概化　进　了比较。研究发现□□□评卷　评分的　□误差　于　□□　的。

实□中,　□种的□□,　测　的评卷　一□　经过比较　□的□　,　式阅卷　会进　比较　□的评卷□□□,　而　研究发现的评卷　评分经　、□□学□□的差　可能　实际的大□度、□□□　文阅卷□□　□□会　□□　试　文　的可□　性。　□　的是 Johnson 及其同□对□□发　评卷　评分差　时　同　理方　的　果的研究。据　的发现,以□论一　的最　评分和通过□□计算　分　为最　评分对　可□　性□　的□□　较大差□。　的研究可□　方向　开,□□发现与　一　,□大可□用计算　分的□□评分方　。

(三)　于评分

评分　是□　□　分　性评分同　也会□□　测　信度。□□(2006)的□□学□论文□□了 40　试 HSK 的　及评分　据,对　同的　□□(　文　论文　□)、评分方(　□　评分　分　性评分)、评分项目(　结　□□□用)　□□　文分　□　各　□的应及　的　□　进　了实□研究。研究发现　□□和评卷　□　应　大　度　评分方　和评分项目□□。□□□　(2008)对 36　□□□□大学　□□　文评分结果的　量进　了分　,□实分项□项评分结果的可□性□于□　评分结果,分项　合分　的可□性□于分项□项分　的可□性,同时□　了□□的信度　对于评分结果的总　量是一　恰当的　。□　的,李智(2009)也从　同　□□和评分　式　发,用概化理论分　了　4　评卷　评的 30　大学　□□分□测试的　文(一　□　、一　题　文)的分项　,□为□　而□,分　性评分　的　有较□的信度。

□□、□□□(2008)的研究□　进一步比较了　分　性评分量　五　评分　□的评分信度,结果　:　、章结　、词汇　用这三　□　的评分较为一　,　对而□,对□□的评分信度最□,而对书　规范的评阅有□提□。□□的研究□向　□　了□□与　□□评卷　对分　性评分　的理　与　用□□的　同:　评分　的理　与　用方　,□□评卷　□重　文的□　、词汇　方　,而　□□　评卷　□　重　文　和连贯性　。

□合这　研究发现,　结合当　的　测评实□,　有　多理　分　性评分　来评试的　能　,这　为它能□来较□的　信度,而□它　提供的评分信息可以　用于　评分实□,从而□低对评分　的主　度,□　评卷　信度,　最　进一步提　可□　性的目　。

(□)其　□□　□

评卷□□分　是　年□　学　开　□的□□　可□　性的又一　□。针对 115　试　□　文和　□　文的　据,Gebril (2010)　的　一项分　结果发现:□　同□　一　可□□评卷　种题型的　文(□　□　)与□　评卷　给　种题型　分　的可□性也　常　。　于　方　的研究　量还　常有　,　有赖于　对　问题多的　究,以期　对□　、　的理　和□　。

外,较之对题目　的　方差的□度重　而□,　试本　的一　征□　的　的方差　的□　当有　,□□中于第　□□　外□□　。　:Solano - Flores 和 Li (2008)通过应用概化理论发现对于□□学□　而□,　试、　文题以及题目本　的□□(□□　外□)的　用是　最大的方差源。Huang (2008)用概化理论研究了　拿大 ESL　试　省级□□　试　的方差源及信度。三年的　据分　结果发现:ESL 和□□为□□的　试的分□　同。ESL　试　文　的残余方差　□比□□为□□的要□。　一年 ESL　试能　释的方差　于□□为□□　试的,其 ESL　试　的概化　低于□□为□□　试

。□文 □ □疑 ESL □ 公□ 。
□ □ □ □ □□ 启□。Gao
□ □ Brennan(2001) 文 □□□若□年 □
□ □ □ 稳 。 □□ □年
□, 且G □ D□策 □□兑 。
□ □ 令□ □ 弃 □带 诸□便 。Gao
Brennan(2001) , □□ ,□ □ 推□ □
情 。□□建□ □□□□变 情□□,尽 □ □ □ ,
□ □ □ 。

四、结 论

言 蔽 , □□□ 映,□□□□ □□ 诸□□
素带 影响, 推□ □□ 思 ,采 □
□ □□ 措□ 保证 靠 。
纵□教育 , □ □ / □十年□ 客□ ,
年□日益凸□ □ / □□□。 □ □/ □归绝□ 单
返□□, 教育 □□□ □, □ / □
推□ 保 , 日益□ □□□ 螺旋推 □归。 □ ,
古 □ □ □ □ □/ , □ 且 □ □
□ □/ 靠 田。 凭 □□ □
长 □ 秘笈, □ 且 □ □ □/ □ 器。 文
□ ,□□ 20 年 □、 □ 推□ 、 文
文献,期望 □ □□ , 日 借鉴 便 。

□□□□:

:《 □ □ □ HSK 高□ 文 □ □ 》, 语言 文,2005 年。
何莲珍、闵□超:《 证 □ 》,《 外语》,2008 年第 6 期。
李 智:《 语 文□□ □ □ □□》,《湖 文 报》(社 版),2009 年第 2 期。
:《□ □□ 文□□ 变□》, 语言 文,2006 年。
□ 、 :《 文 □ 》,《□ 报》,1998 年第 2 期。
:《 □ 语 □ □□ 》,《 》2008 年第 5 期。
乔治·恩舟□□德:《 □□ Rasch □ 》,□□译,《教育 》,2007 年第 4 期。
、祁宗海、席仲恩:《 文整 □ □□ □ □ 》,《外语 》,2008 年第 5 期。
:《HSK □ □ —— □ 调 》,《 》,2010 年第 10 期。

Bachman, L. *Fundamental considerations in language testing*. Shanghai: Shanghai Foreign Language Education Press, 1999.

Engelhard, Jr., G. Differential Rater Functioning. *Rasch Measurement Transactions*, 2008, 21(3).

Gao, X. & Brennan, R. L. Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, 2001, 14(2).

Gebril, A. Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 2010, 15(2).

Gebril, A. Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, 2009, 26(4).

Huang, J. How accurate are ESL students´holistic writing scores on large – scale assessments? ——A generalizability theory approach. *Assessing Writing*, 2008, 13(3).

Johanson, G. & Alsmadi, A. (1998). *Differential Person Functioning.* ED 420 691.

Johnson, R., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly*, 2005, 2(2).

Lee, Y.-W. & Kantor, R. Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing*, 2007, 7(4).

Linacre, J. M. *Many-facet Rasch measurement.* Chicago, IL: MESA Press, 1989.

Nie, Y., Yeo, S. M. & Lau, S. Application of generalizability theory in the investigation of the quality of journal writing in mathematics. *Studies in Educational Evaluation*, 2007, 33(3 – 4).

Rasch, G. *Probabilistic models for some intelligence and attainment tests* (revised and expanded ed.). Chicago: The University of Chicago Press, 1980.

Schoonen, R. Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 2005, 22(1).

Sim, J. & Wright, C. *Research in health care: concepts, designs and methods.* Cheltenham, UK: Stanley Thornes, 2000. http://books.google.com/books?id=vwjhgtUoNZIC&pg=PA335&lpg=PA335&dq=%22estimate+of+Intra-rater+reliability%22&source=bl&ots=6FQUcpr6X5&sig=fY5VEyJ_BqG54wSv0w61GVazYAM&hl=en&ei=It7LTIH2EYmAvgO0iZnMDw&sa=X&oi=book_result&ct=result&resnum=1&ved=0CBIQ6AEwAA#v=onepage&q=%22estimate%20of%20Intra-rater%20reliability%22&f=false. 2010 – 10 – 30.

Solano-Flores, G. & Li, M. Examining the dependability of academic achievement measures for English language learners. *Assessment for Effective Intervention*, 2008, 33(3).

Sudweeks, R. R., Reeve, S. & Bradshaw, W. S. A comparison of generalizability theory and many – facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 2004, 9(3).

Wikipedia. *Intraclass correlation.* http://www.answers.com/topic/intraclass – correlation. 2010 – 10 – 30.

The Developmental Trend of Open-Ended Assessment Based on Reliability Research of Writing Tests: A Perspective of Generalizability Theory

FUNG Shui-Lung & ZHU Yu

(Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai 519085, China;
Overseas Education College, Xiamen University, Xiamen 361102, China)

Abstract: Subjective test has a long history, a strong vitality, and a prosperous future. Meanwhile, its low reliability, considered as its 'chronic disease', has been disturbing practitioners and researchers in the field of educational measurement. This paper reviewed the three major schools of reliability theory, and clarified that Generalizability Theory is a forward – looking theory which can offer approaches of improvements based on estimates of variance components of various effects potentially lowering the reliability of a test. Implications to the development of subjective tests were briefly discussed based on a literature review of writing assessment studies from the perspective of Generalizability Theory.

Key words: subjective test; reliability; Generalizability Theory; writing study