25 期活动讨论主题：评分人经验与考试评分

评分人差异或者评分人效应是语言测试研究的热点，一直受到国内外学者的广泛关注，其中评分经验是影响评分质量的关键因素。国外对此有不少针对性研究，但是国内的研究缺乏针对性，且较零散。这篇文章的作者徐鹰老师采用混合研究设计，既对分数进行定量统计分析，又让评分人写评分理由，然后对评分理由进行量化分析，力求对不同经验的评分人差异进行深入探索。

点评文献：

徐鹰，2015，评分经验对 CET-4 作文评分人差异的影响研究[J]，《中国外语教育》，（03）：74-84。

论坛嘉宾：

中国科学院大学孙海洋副教授

华南理工大学徐鹰老师

思考题回答

1. 作者是如何引出文章的研究问题的？

孙海洋老师：在引言部分作者首先综述了国外和国内有关评分经验研究的现状和以往的研究发现、所采用的研究设计和方法，随后指出评分经验研究对我国大规模考试中写作评分工作的重要意义，在此基础上自然引出自己的两个研究问题。两个问题可以概括为：不同评分经验评分人的评分结果是否存在差异，他们所给的评分理由是否存在差异。

2. 作文文本是通过什么方法选取的？这种选择方法有什么优越性？

孙海洋老师：作者通过随堂作文收集 CET-4 作文文本，再对全部作文按照 CET－4 评分标准进行初评，并根据初评结果采用分层随机抽样的方法抽取了 30 篇涵盖 5 个分数档的作文作为研究材料。

采用这种方法收集的文本涵盖了所有层次水平，控制了每个层次的样本数，使之与现实状况基本相同，比较全面且具有一定的代表性；同时又控制了样本的总数，使后面的研究中评分工作量不会太大，可操作性强。

3. 作者是怎样对研究被试进行分类的？是用什么统计方法检验三组评分人评分次数差异性的？

孙海洋老师：作者按照评分次数将被试分为 3 类：新手（1≤评分次数≤5）、中手（6≤评分次数≤10）和老手（评分次数≥11）。研究采用单因素方差分析对比三组评分人的平均评分次数有无差异。由于样本量较小（每组仅有 9 个人），最好采用 Kruskal Wallis 非参数检验方法来检验三组被试的评分次数平均值是否有显著差异。

4. 文中所给的研究的自变量和因变量各是哪些？这些变量分别是什么类型的变量？是针对哪个研究问题的变量？

孙海洋老师：从研究问题来看的话，第一个问题的自变量是评分经验，是称名变量（nominal variable），因变量是评分严厉度、一致性和随机效应，但是在测试研究中，这些都是评分信度分析，属于测试指标的描述性统计量，一般不区分自变量和因变量。

第二个研究问题包括两个自变量：评分经验，为称名变量（nominal variable）；和评分理由重要性，是连续变量（continuous variable）；5 个因变量分别为切题、表达思想清楚程度、连贯、语言错误以及篇幅等编码频数占全部评分理由编码频数的百分比，这 5 个变量都是比率变量(ratio variable)，也是一种连续变量(continuous variable)。

5. 数据分析采用什么方法？为什么要用这些方法？

孙海洋老师：数据分析分别采用多层面 Rasch 模型和混合多元方差分析 MANOVA 方法，用 FACETS 和 SPSS 软件来实现。测试领域对于评分一致性的研究传统上采用相关分析的方法，但相关分析只能考查评分人之间的一致性，而个体评分者自身前后的一致性无法通过相关分析的方法得知，多层面 Rasch 模型既可以检测评分者之间的一致性，也可以检测评分者内在的一致性，因此，回答第一个研究问题作者采用了这种分析方法。

第二个研究问题的因变量通过编码计算百分比之后都成为了连续变量，且相互之间有一定关系，为了有效控制并减少犯第一类错误的概率，作者没有采用多个 ANOVA 的方法，而是直接采用了 MANOVA，其结果要比多次 ANOVA 的结果更为可信，也比不做编码百分比计算只做单纯的频次卡方检验更为细致、可靠。

6. 作者是怎样回答第一个研究问题的？FACETS 分析结果中哪些值可以体现评分人的内部一致性？哪些值可以体现评分人之间的一致性？logit 值的大小如何对应评分的严厉程度？

孙海洋老师：对于第一个问题的回答作者采用了多层面 Rasch 模型的方法，首先从个体层面上分析了 27 个被试的评分一致性，指出这些评分人表现出明显不同的严厉度，内在一致性也存在差异；然后又从小组层面上比较了 3 组评分人的评分一致性，指出整体上三个小组并没有出现明显不同。FACETS 分析结果中加权均方拟合值（Infit MnSq）体现评分人的内在一致性。作者将参考值设定在 0.7 到 1.3 之间，大于 1.3 的评分人评分不拟合（misfit），前后一致性比较差，时而严厉时而宽松；小于 0.7 的评分人评分则过度拟合（overfit），评分过于集中，没有区分考生的差异。FACETS 结果中的分隔信度和卡方检验结果体现评分人之间的一致性，分隔信度越接近 1，卡方检验结果越显著，那么评分人之间的严厉度差异越大。严厉度 logit 值越小（负值）说明评分人越宽松，越大越严厉。

7. 作者是如何回答第二个研究问题的？评分理由是通过什么方法量化的？MANOVA 分析结果说明了什么问题？

孙海洋老师：第二个问题作者采用混合多元方差分析的方法来回答。评分理由重要性是根据被试所列三条理由的排序来量化的，具体的评分理由是通过计算其频次占总频次的百分比来量化成连续变量的。MANOVA 的分析结果显示，评分经验主效

应不显著，评分经验和评分理由重要性的交互效应也不显著，但是评分理由重要性的主效应显著，但仅在切题和语言错误上存在显著差异。事后分析显示，切题在第一条评分理由上的百分比显著高于其在第二条和第三条评分理由上的百分比，语言错误在第一条评分理由上的百分比显著低于其在第二条和第三条评分理由上的百分比。这些结果说明，不同经验的评分人都会首先评判作文是否切题，然后观察作文的语言错误，从而作出评分决策。

8. 这篇文章的两个研究问题的内在联系是什么？评分标准表征对于提高评分信度有无帮助？

孙海洋老师：第一个问题是表象，是不同评分经验的评分人表现出来的评分结果差异（尽管组间没差异，但个体有差异）；第二个问题是第一个问题的延伸，看第一个问题里的评分差异是否会在评分理由上有所体现，尽管结果是组间（不同评分经验小组之间）没有差异。

关于评分标准表征和信度的关系可以从两个方面来看：一方面，对于采用整体评分方法的考试，太多的评分标准表征不一定有利于评分信度的提高，反而是单一或者相对集中的评分标准表征对提高评分信度有帮助；另一方面，不同的评分标准在整体分数里所占的比例应该是不同的，比如这项研究中“表达思想清晰程度、连贯和篇幅”占评分理由总数百分比比较低的原因可能就是评分老师们认为这些标准“不重要”，不足以影响考生作文的整体质量。若要体现这些标准的重要性，一是可以采用分项评分方法，将这些评分标准作为相应的评分维度；再就是根据需要对分项分所占整体分的比例做出合理、明确的规定。

关于评分标准表征，徐鹰老师补充：个人感觉评分标准表征是一个效度问题，涉及到分数的意义和可解释性。不同评分人给出的相同分数的意义可能完全不同，因此威胁了分数的效度。我们希望评分人打分能够从评分标准出发，覆盖全面的评分标准相关特征，这样就能保证分数的效度。但是，由于要考虑多个特征，如何整合（integrate）这些特征并得出最后的分数是一个复杂的心理决策过程，需要有一个定义清晰、描述准确、层次有区分性的评分标准作为保障，否则就容易造成评分人判断混乱，从而影响评分信度。

9. 研究结果对于大规模写作测试评分有哪些启示？

孙海洋老师：除了作者在文中提到的两点，我个人感觉还有以下两点启示：第一，尽管不同经验小组之间的评分严厉程度并不存在显著差异，但对评分人个体而言，不同的评分人之间还是存在显著的严厉度差异的，在大型考试评分中这些差异某种程度上可以通过双评得到消除或缓解；然而评分人自身的内在一致性却会严重影响评分的信度和可靠性，那么大规模考试的评分培训应主要针对提高评分人的内在一致性，或者说要重点培训内在一致性存在问题的评分人。

第二，关于评分标准表征的问题，尽管整体评分的方法容易导致评分还原主义（reductionism）倾向，然而现实中分项评分往往不可行，只能进行整体评分。如果在整体评分标准里能对各项标准孰重孰轻有明确的排序，或者有比较明确的规定或提示，那么评分老师在评分过程中才有可能会兼顾各项标准所占总分的比例。

关于第二点，徐鹰老师补充：这一点其实也是未来研究的方向之一。评分标准的构

建是目前国际语言测试界研究的热点，希望能引发更多国内学者的关注。

互动问答

1. 文章中评分人只有 27 位，分组的样本数更少，是否会有基数不够大，这样利用统计工具得出的数据是否不够具有代表性和普遍性的质疑呢？

   徐鹰老师：这个问题是关于样本量的。一般评分研究的评分人样本都比较小，因为现实中很难找到那么多老师同时来评分，无法像教学研究那样找到大量学生。在这篇文章中，27 位评分人相比同类文章的样本数目较多。本研究的统计是对评分人提供的评分理由进行 MANOVA 分析，有 2196 条理由，这个量是相当大的。另外，对评分人的严厉度、一致性进行 MFRM 分析，MFRM 分析的亮点就是实现放大镜的功能，可以对每个评分人的特点进行细致分析。

2. 在表 6 中，评分理由按重要性列出了五项（具体见论文），而我本人阅卷时有一个比较大的困惑，是这五项不能代表的，想知道培训时老师会提到吗？这个困惑就是学生经过培训机构的模版教育，写出来的句子是一套一套的，语言本身的确错误少，但读起来很别扭。这种情况评分人如何处理呢？

   徐鹰老师：我想这个和 CET-4 作文培训有关。广州阅卷点特别要求老师从评分标准规定的文本特征出发来评分。尽管在实际操作中有老师会简化为只看切题、语言错误等特征。此外，我们也经常发现您谈到的模板作文情况。这个是目前国际上对中国考生作文水平评价最头疼的问题。就这个问题，可以进行深入探讨。

3. 如何对评分员进行有效的培训？
   徐鹰老师：评分员培训也是很多专家提到过的，我的博士论文《评分人培训的研究现状及展望》做的就是这方面。
   孙海洋老师：大型考试阅卷的老师应该都有过被培训的经历，简单来讲就是这样几个环节：挑样卷，专家给样卷打分，培训老师试评样卷，比对专家分，再试评，直到标准统一。

4. 国外的考试，如 SAT，有人研究过其中的评分人评分差异吗？
   徐鹰老师：这方面研究比较多，评分人差异是语言测试的研究重点。大家可以关注几个主要作者，Barkaoui （2010）、Wiseman（2012）， Yan （2014）， Eckes （2008，2010）发表在 Language Testing , Assessing Writing 和 Language Assessment Quarterly 上面的文章

5. 我对于文章前半部分用的 Facets 还算熟悉，但是后边提到的混合多元方差分析不了解，不知能否普及下？
   徐鹰老师：混合多元方差分析是参数检验，数据必须是连续性数据，不能用频数，这个统计方法功能很强大，并且能控制一类错的概率。

6. 虽然经过培训，但是否有时评分员很难保持自身一致性？ 评分员宽严度也有差异？既然 facets 可以进行检验这些及评分偏差，那么是否在大型考试评判过程中通过软件进行监控及时调控呢？

徐鹰老师：大规模考试比较难实现 facets 的实时监控。其实主观题的评分很值得研究。桂诗春老师举过一个例子。1983 年北京师范大学张厚璨教授在高考评分前从北京随机抽取语文、政治、数学、物理每科五份考卷, 复印后分发到全国除西藏、台湾的 28 个省、市自治区, 请各地高考阅卷组分头评分, 结果发现同一份试卷误差程度很大, 评分误差最大的是语文, 政治次之。语文科中, 同一份试卷最大差异达 33 分, 最低有 19 分, 说明各地区之间存在着较大的评分误差。

7. 关于评分理由编码，能否再详细介绍下编码如何呈现？ 如用数字赋予不同的理由？
徐鹰老师：定性数据的分析可以参考 Miles & Huberman （1994）的专著，国内定性研究可以参考北大陈向明老师的专著和论文。

8. 表 3 评分理由编码框架里，有“长度”这个因素，看字数有没有达标，没达标是指字数少了吗？如果字数超出很多，是否也算没达标？
徐鹰老师：长度是指作文篇幅，四级评分标准有提到这一点，同时评分人的评语也有提到，因此必须进行分析。关于作文词数的规定，我个人认为有必要做一些实证研究，给出具体词数的参考依据。目前要求学生写多少词感觉都是老师主观判定的，缺乏实证依据。在我的研究中，出现问题的都是字数不达标的，因此只用这个就行了

孙海洋老师：命题作文基本是这样，但是像概要写作和自由写作考察的概念不一样，应该要有篇幅限制，字数多的话也要扣分的。四六级作文题目中有明确规定，少于 120 字扣分。

9. 国内有哪些考试实行机考了？
徐鹰老师：最著名的就是 TOEFL ibt，CBT。另外，剑桥的博思考试也有计算机化考试，广东省高考英语听说考试就是机考。
孙海洋老师：四六级考试现在有网考。