[中图分类号] H319　　[文献标识码] A　　[文章编号] 2095-1167(2018)01-0012-07

以“听读后写”为特点的医务英语写作测试评分量表研究*

贾文峰[1]　　张培欣[2]

提要： 语言测试中的行为测试常和专门用途英语测试相结合。本研究设计了一项模拟现实交际任务的医务英语写作行为测试：考生扮演医生的角色并完成对某位患者的诊断过程，在获得相关信息的基础上撰写英文病历书，测试形式以“听读后写”为特点，测试具有真实性。评分量表兼顾语言技能和基于特定工作的行业技能，体现现实交际任务的评价准则。三个评分分项分别为语言运用水平、对病情的理解和把握情况、病历书的结构和逻辑。评分结果分析表明，该量表具有较好的信度和效度。该测试的任务形式和评分量表都独具特色，对类似的专门用途英语测试的开发和研究具有一定的借鉴意义。

关键词： 行为测试；医务英语写作测试；评分量表；信度与效度

Abstract: Language performance test is usually associated with assessing of English for Specific Purposes. In this study, a medical English writing test is designed, in which the candidates are required to take the roles of doctors, to complete the process of diagnosis and then to write the formal case records based on the information about the patient. This test is of authenticity and features the format of writing after listening and reading. The rating scale of this test takes the elements of language proficiency, professional skills and assessing criteria for doctors´ jobs in reality into consideration. Three dimensions, language performance proficiency, understanding and mastery of the patient´s condition and the structure and logic of case record, are included in the rating scales. The analysis of the rating results shows that the rating scale has good validity and acceptable reliability. The test, characterized by its distinctive test format and rating scale, provides reference for designing and validating similar tests of English for Specific Purposes.

Key words: performance test; medical English writing test; rating scale; validity and reliability

1. 引言

语言测试中的行为测试(performance test)也称为真实测试或直接测试，以现实生活中的语言运用行为为准则，努力使测量过程接近非测试环境中的语言运用行为(Davies *et al.* 1999)。行为测试经常和专门用途英语测试结合在一起，通常用来评价应试者完成与其工作或学习相关的具体任务的能力，强调测试任务与现实语言使用目标领域的对应或吻合，这类测试的设计往往建立在工作分析的基础之上(McNamara 1996; Douglas 2000)。换言之，与专门用途英语测试相结合的行为测试，其突出特点表现在测试形式或测试方法上。测试的任务情景接近或模拟现实中语言使用情景；应试者在测试任务中的角色往往就是现实工作中的角色；测试任务的语言输入和输出特征接近现实交际任务中的情况。

* 本文为 2016 年教育部人文社会科学研究青年基金项目“新 TEM-8 写作测试评分量表建构研究”(编号 16YJC740029)阶段性成果之一，“第四届全国外语测试学术研讨会”宣读论文。

语言行为测试常采用以“说”或“写”为语言运用方式的测试任务。由我国民航局批准的民航管制人员英语等级考试(Air-traffic Controller English Test System,简称 AETS)就属于这类测试,其第四部分即“模拟通话”的任务中,要求应试者以管制员的角色在模拟的国际航空运行场景中运用英语完成各种交际任务。无论是“说”还是“写”的测试任务,都需要借助评分量表对任务进行评价或评分。以写作为例,评分量表有整体和分项评分量表等基本形式。分项评分量表的不同分项可以称为不同的准则(criteria),每条标准包括不同级别(ranks)以及相对应的描述语(descriptors),评分人要对样本基于不同的准则分别评分(Weigle 2002; Knoch 2009)。基于专门用途英语的语言行为测试中,考生要完成测试任务不仅需要具备语言能力,还需要有专门的学科或行业知识;评分量表的准则及其描述语不仅涉及语言因素本身,还要参照现实专门行业中交际任务完成情况的评价准则(Bachman & Palmer 2010),与其它语言测试形式相比,这类测试从任务形式到评分量表都具有自身一定的特点。

2. 测试任务特色与研究问题

笔者团队在对某大型涉外医院调研后,发现用英文写完整病历是常见的写作任务。该院已与超过 30 家国际保险公司建立跨国结算业务,外籍病人在该院就诊的费用可直接由外国保险公司支付,中方医生用英文撰写符合规范的病历书是工作必需,也是国际结算的凭证。在进一步了解相关情况之后,以院方专家所提供的一件由其本人撰写的真实英文病历为母本,笔者团队设计了一项医务英语写作行为测试(写作提示语见附件),设计原则是将现实涉外医务工作场景中病历产生过程尽量对应到测试任务中来。现实医务中,医生的问诊过程包括询问主诉与现病史、询问既往病史社会史与家族史等,在测试任务中也模拟患者和医生之间基于这些问诊过程的英语对话,分别由一名英籍外教和中方教师扮演写作题目中 Susan 和 Dr. Li Ming 的角色并录音。医生和患者对话共计1 488个英文单词、音频长度为 12 分钟。现实医务中,医生还需要阅读来自其他医生的转诊信息并阅读相关的检查结果,在医务英语写作测试的材料中也包括相应的转诊信息和检查结果,附有血常规和心电图检查结果等阅读材料。因此,从测试任务的输入和输出特点来看,该项考试可以称作是以“听读后写”为特色的医务英语写作测试。考生的写作时间为 50 分钟。

本文的研究问题是:该项专门用途英语测试中,评分量表如何既体现英语语言能力,又体现医学学科知识及专业技能决定的现实交际任务的完成情况?此种评分量表的效度和信度如何?

3. 研究步骤、方法与工具

3.1 施测与样本收集

该测试的实施对象是某 985 重点大学的医学院临床医学专业七年制本硕连读大三学生。这些学生正在选修《医务英语阅读与写作》,英文病历写作是课程内容的一部分。将该写作测试任务作为学生期中考试,目的是考查学生课程学习情况和与之相关的专业英语写作水平。共收集到写作文本 44 份。

3.2 界定评分分项所测能力及名称

在没有量表的情况下通读样本,基于研究者的主观判断挑选出能够代表“高”、“中”、“低”三个层次的样本各一份,参照样本特征确定评分分项描述语及级别。关于量表的级别,一般而言分为 5 个或 6 个级别都比较合适(Weigle 2002)。在分项评分量表确定后,聘请 4 位医务英语写作课的授课教师依据评分量表对样本进行评分。

3.3 验证量表的效度和信度

评分量表的效度可以从多方面论证。在本研究中，量表的效度指：量表的评分结果或分数展现出的规律对量表的分项或标准所测能力的界定的支持程度；信度则指评分人间的一致性（Knoch 2009; Brown 2012）。主要利用的统计工具包括 Rasch 分析软件以及进行相关分析和信度系数分析的 SPSS 软件等。

4. 研究结果与讨论

4.1 评分分项所测量的能力界定以及量表的内容

根据前文的讨论，该量表的分项要兼顾语言因素和现实专业交际任务要求。语言因素主要涉及信息组织能力。听力材料中的医患对话体现了面对面的口语交际特点，医生在交际中居于主导地位，对话中存在停顿、犹豫、意义协商等，冗余信息较多；书面阅读材料转诊书和化验单中的信息也多以短语形式呈现，需快速提取信息；而书面病历的英语语法则体现书面语的特征。例如，围绕主诉部分，听力材料用 16 轮一问一答的形式来交代，而写作时可以将“疼痛的名称、位置、持续时间、方式、时间、频次”等信息凝缩在一个英语复合句中。因此量表的第一个分项命名为“语言运用水平”，简称“语言”分项，主要指考生要具备相当的语言组织能力或语法能力。

第二个评分分项是“对病情的理解和把握情况”，简称“理解”分项，主要指考生对写作输入材料（包括听和读的材料）中所包含的近 40 个关于病情的信息点的理解能力。

第三个评分分项是“病历书的结构和逻辑”，简称“结构”分项，主要是将信息组织成书面病历的结构安排能力。工作中书面病历要求按照一定的格式来安排各个部分，不同部分的内容要有逻辑性和一致性。

确定以上准则或分项之后，参照所选取的能代表高、中、低水平的样本确定 5 分档、3 分档和 1 分档描述语，各个分项设 5 个级别，所制定的评分量表如表 1 所示：

表 1　医务英语写作行为测试评分量表

	语言分项	理解分项	结构分项
5 分	复合句、并列句和简单句交替运用，符合书面病历特点； 时态运用具有丰富性； 熟练运用介词短语、分词等形式将信息组织成书面语，体现了和口语不同的语法特征	对案例中病人病情有较好的理解和把握； 正确理解绝大部分信息点特别是关键信息，仅遗漏个别信息	病历书各部分结构安排合理； 各部分之间的逻辑性强
4 分	介于两者之间	介于两者之间	介于两者之间
3 分	样本有一些语法错误，时态运用具有一定丰富性； 可以运用介词短语、分词等形式将信息组织成书面语，但是有欠缺	基本把握病人的病情； 部分重要信息点理解不够准确和全面，但不影响诊断	结构安排基本合理； 有些地方逻辑性欠缺
2 分	介于两者之间	介于两者之间	介于两者之间
1 分	绝大部分句子存在语法错误； 基本都是简单句；力图使用复杂句，但基本都有错误； 动词的各种时态形式运用、介词短语、书面病历常用句型等使用贫乏、有限	理解非常有限； 不能把握病人的基本病情，不能进行正确的诊断	结构安排混乱； 逻辑性差且前后矛盾

由表 1 可见，考生要完成写作任务，除语言能力之外，还需要利用自己的医学专业技能和专业推理能力来把握病情并做出诊断、组织信息并合理安排病历书的结构，这些超出了语言因素本身。量表所界定的三个方面的能力有所区别，体现了对语言因素和专门行业知识和能力的兼顾。取三个分项成绩的平均分作为最终的写作成绩。在阅卷开始前，对评分人进行适当培训，使他们充分了解测试的任务特征及能力要求，熟悉量表的描述语和使用方法。最后收集四位评分人的评分结果用于分析。

4.2 评分量表的结构效度验证结果

各个评分分项的成绩，可以看做是基于同一项测试任务不同测试部分的成绩，不同测试部分所测量的能力各有所侧重(Bachman 1990) 。对于同一测试的不同部分而言，彼此之间的相关系数应该是既不太高又不太低，相关系数在 0.3 到 0.7 之间最为合适，这说明各个部分考查的具体能力既有相关性，又存在差异性(Alderson 1995) 。本研究中的写作测试的评分分项可以看做是基于同一写作任务测量不同能力的三种测试形式。第一是通过“听读后写”的方式，测量考生将理解到的英语信息组织成恰当的语篇的能力; 第二测量考生用英语理解和把握病人病情的能力，包括信息的正确性和全面性; 第三是测量考生将信息组织成书面病历的结构安排和逻辑能力。四位不同的评分人的评分结果，可以理解为同一项测试不同的平行测试的情况。基于这个视角，该评分量表的三个评分分项的相关关系也应该在合理区间之内。

表 2 反映的是四位评分人的各个分项评分结果之间的相关系数。从表 2 可见，在同一项测试的四种平行测试中，除评分人 1 的语言与理解、评分人 2 的理解与结构以及评分人 4 的语言和理解之间的相关系数稍微高于 0.7 之外，其它分项之间的相关系数均在 0.3 到 0.7 之间，这就从统计结果上证明了上述假设，三个评分分项所体现的能力之间既有关系又存在区别性，从这三个方面进行评分是合理的。

表 2 各评分人各个分项分数的相关关系

评分人 / 相关系数	1	2	3	4
语言与理解	0.703**	0.347*	0.687**	0.736**
语言与结构	0.490**	0.562**	0.443**	0.646**
理解与结构	0.581**	0.707**	0.639**	0.688**

** 表示在 0.01 的显著性水平上具有显著性。

* 表示在 0.05 的显著性水平上具有显著性。

分项评分量表的效度还可以从维度效度(dimension validity) 的角度进行举证。维度效度指分项评分量表的各个标准的区别性，或者各个标准是否很好地反映了写作能力的不同方面(Bachman 1990; McNamara 1996) 。通常，在采用多面 Rasch 分析方法的过程中，将每个评分分项或每个标准看作不同题目，而题目中的每个等次看作不同范畴。如果某个标准或分项的评分结果出现了非拟合(misfit) 的情况，说明这个标准的设计和其它标准相比不够合理或区分度不高，例如，能力高的应试者可能在此项得分低，而能力低的应试者在此项可能得分高。如果某个标准的结果出现过度拟合(overfit) 的情况，说明单用这个标准就可以很好地区分应试者，或者说，分别评价各个标准并取均分的做法得到的成绩，和单独利用这个标准进行评分的效果

一致。如果语言这个评分分项出现过度拟合，就说明用这一个评分分项评出来的成绩和利用再加入其它分项的量表评出来的成绩，在区分应试者或评测效果上是完全一致的，这就削弱了采用不同分项进行评分的价值和必要性(McNamara 1996) 。因此，理想的情况是，各个分项既不出现非拟合也不出现过度拟合的情况。表 3 是对评分结果进行基于多层面 Rasch 分析的相关结果。

表 3　基于 Rasch 分析的分项量表各标准的统计量

评分标准	难易度	标准误	加权均方拟合值	标准 Z 分数
语言	0.09	0.16	0.95	−0.4
理解	−0.08	0.15	1.03	0.2
结构	0.00	0.15	0.86	−1.2

上述结果中，判断题目拟合情况的数值是加权均方拟合值和标准 Z 分数。一般而言，加权均方拟合值接近 1 而且标准 Z 分数的值在正负 2 之间说明题目的拟合情况良好(Linacre 2005; 刘建达 2005) 。如果加权均方拟合值超过 1.3 而且标准 Z 分数的值大于 2，说明题目属于非拟合的情况; 如果加权均方拟合值低于 0.75 而且标准 Z 分数的值小于−2，说明题目属于过度拟合情况。从表 3 看，语言、理解和结构三个分项的加权均方拟合值分别为 0.95，1.03，0.86，都比较接近 1，没有超出 0.75 至 1.3 的范围，而且各自的标准 Z 分数的值也在正负 2 之间，拟合情况均比较理想。这说明各个评分分项有效地反映了写作不同方面的质量，从不同角度各自独立地表征了写作能力。单独使用语言分项、理解分项或结构分项中的任一单个标准，都不能全面反映考生要完成此项写作行为测试任务所需具备的能力。这进一步证明，此类考试评分量表的内容要同时体现专门用途语言技能和专门行业技能，兼顾语言要素和现实交际任务完成情况评价准则，仅仅使用语言能力一个分项或维度进行评分是不全面的，需兼顾各个分项，将分数相加，使分项评分量表具备较好的维度效度。

4.3 评分结果的信度分析

表 4 是分项评分量表各分项的评分信度系数以及分项均分的评分信度系数。本文的量表信度水平指评分人间的评分信度，需要考虑不同的评分人的组合情况(Brown 2012) 。“1&2”表示评分人 1 和评分人 2 之间的评分信度，“1&2&3”则表示评分人 1、评分人 2 和评分人 3 之间的评分信度，其余类同。在评分人组合 1&2 中，语言、理解、结构分项的评分信度系数分别为 0.78，0.71 和 0.80，而分项均分的信度系数为 0.87，分项均分的评分信度高于分项评分信度。直观上来看，分项均分的评分信度无论在两人组合还是三人组合中，都高于各个分项的评分信度。从表 4 中可以看出，三人分项评分的均分信度系数均在 0.90 以上，这说明在专门用途英语写作行为测试中，兼顾语言要素和现实行业任务完成情况准则的评分量表，完全可以达到理想的信度水平。

表 4　分项评分和整体评分的信度比较

组合	语言分项	理解分项	结构分项	分项均分
1&2	.78	.71	.80	.87
1&3	.81	.78	.75	.90

（续表）

1&4	.88	.75	.87	.89
2&3	.75	.69	.81	.85
2&4	.77	.69	.74	.80
3&4	.85	.89	.79	.93
1&2&3	.85	.81	.84	.91
1&2&4	.87	.79	.87	.90
2&3&4	.86	.84	.84	.91
1&3&4	.89	.87	.87	.94
1&2&3&4	.90	.87	.89	.94

进一步采用方差分析的方法比较表 4 中四列信度系数的差异性。表 5 是方差分析结果，该结果说明在满足方差齐性的条件下，语言分项、理解分项、结构分项和分项均分四个方面的信度系数存在显著性差异（$F(3, 40) = 6.736, p<0.05$）。方差多重比较结果发现，分项均分的评分信度显著高于理解和结构分项的评分信度，但和语言分项的评分信度没有显著性差异；同时发现语言分项的评分信度和理解与结构分项的评分信度没有显著性差异，虽然语言分项的信度平均数为0.84，略高于理解分项的信度平均数 0.79 和结构分项的信度平均数 0.82。这说明，不同的评分分项虽然对语言能力本身和行业技能的侧重点不同，但都可以取得较为理想的评分信度。总体而言，本部分基于评分结果的分析表明，该医务英语写作行为测试的评分量表具备较好的信度，能够运用于评分实践。

表 5　各分项及分项均分的评分信度比较

	语言分项（n=11）	理解分项（n=11）	结构分项（n=11）	分项均分（n=11）	F(3,40)
平均数	.8373	.7900	.8245	.8945	
标准差	.05159	.07335	.05067	.04180	6.736

5. 总结与启示

当前，专门用途英语的教学改革和研究不断发展，教学形式和教学内容日益丰富和多样化。在各类各层次的教学机构中，不乏见到诸如导游英语、航海英语、国际金融业务英语等基于特定行业或职场的语言教学活动。如何使测试形式服务于这些教学活动是值得探索的问题。特别是英语教学和用人单位的入职考试结合在一起的情况下，开发此类基于特定行业的语言行为测试就非常必要。这种测试能够直接考查应试者用英语完成专门用途交际任务的情况，测试的分数可以有直接意义，由现实的语言交际任务和测试任务之间的相似性来保证测试的表面效度，用需要分析或工作分析来保证测试的内容效度，测试的使用者可根据应试者的测试表现直接推测其在非测试环境下的语言使用情况（王振亚 2009）。这种测试显然是大学英语四、六级考试等大规模标准化语言测试所不能替代的，两类测试的性质和用途都不尽相同。

然而，当前关于此类测试的研究却非常缺乏。本文在医务英语写作测试的方面进行了一定的探索，旨在抛砖引玉。本文的研究再次表明，工作分析作为此类测试的测试任务设计及量表开发的关键环节，为整个测试提供了实证基础。一方面，通过工作分析可以描述现实行业中用英语完成的交际任务的语境特征，包括语言输入和输出特征、语言使用渠道等，通过将这些现实语言交际任务的要素复制或模拟在测试任务中，保证语言测试任务的真实性。另一方面，只有在和现实交际任务相吻合的测试任务中，分项评分标准才能既反映语言因素，又兼顾现实交际任务完成情况和评价准则。在满足以上条件的基础上，才能保证测试使用者依据应试者在测试中的表现有效地推断其在非测试环境下的语言使用情况。以上推理过程和 Bachman & Palmer(2010) 提出的基于现实世界的测试使用论证框架(Assessment Use Argument，简称 AUA) 实则一致。

希望本研究能引起大家对该类测试更多的关注，根据现实需要开发和研究具备行业特色的测试任务和量表形式，更好地服务于语言教学和测试实践。

参 考 文 献

[1] Alderson J C. *Language Test Construction and Evaluation* [M]. Cambridge: CUP, 1995.

[2] Bachman L F. *Fundamental Considerations in Language Testing* [M]. Oxford: OUP, 1990.

[3] Bachman L F & Palmer A. *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World* [M]. Oxford: OUP, 2010.

[4] Brown J D. *Developing, Using, and Analyzing Rubrics in Language Assessment with Case Studies in Asian and Pacific Languages* [R]. University of Hawaii: National Foreign Language Resource Center, 2012.

[5] Davies A, Brown A, Elder C, Hill K, Lumley T & McNamara T. *Dictionary of Language Testing* [Z]. Cambridge: CUP, 1999.

[6] Douglas D. *Assessing Language for Specific Purpose* [M]. Cambridge: CUP, 2000.

[7] Knoch U. *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale* [M]. Peter Lang GmbH, 2009.

[8] Linacre J M. *A User's Guide to FACETS: Rasch Measurement Computer Program (Computer Program Manual)* [Z]. MESA, Chicago, 2005.

[9] McNamara T F. *Measuring Second Language Performance* [M]. London: Longman, 1996.

[10] Weigle S C. *Assessing Writing* [M]. Cambridge: CUP, 2002.

[11] 刘建达. 话语填充测试方法的多层面 Rasch 模型分析 [J]. 现代外语，2005，(2)：157-169.

[12] 王振亚. 现代语言测试模型 [M]. 保定：河北大学出版社，2009.

附件：医务英语写作测试任务提示语

Suppose that you are Doctor Li Ming. Now you are treating a foreign patient for chest pain. The listening material is about the dialogue between you and your patient, which will cover the necessary information about her. Besides, the written materials such as the referral letter and results of laboratory analysis will be helpful for your diagnosis.

Please try to understand the listening and written materials for a good mastery of the patient's condition. And then, write a formal case record for the patient, which will be used by an international health insurance company.

作者联系方式： 1. 山东大学(威海) 翻译学院，山东 威海 264209

2. 厦门大学外文学院，福建 厦门 361005