心理学报 2011, Vol. 43, No.4, 462−472
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2011.00462

数学应用题中语言成分模型的构建 —— 多元随机效应项目反应理论模型的运用*

杨向东[1] Lorraine Chiu[2] 卫 芸[1]

([1]华东师范大学课程与教学研究所, 上海, 200062) ([2]堪萨斯大学教育心理学系, 美国)

摘 要 数学应用题中的语言成分可能对被试问题解决过程产生复杂影响。通常, 这种影响对所有被试并非完全一致, 而是具体数学题目特征与特定被试的认知特性之间交互作用的结果。本研究采用多元随机效应项目反应理论模型的建模方法, 分析了数学应用题中语言成分对问题解决过程的影响。该方法的优势在于它不仅分析了语言成分对数学问题解决过程的平均效应, 同时给出了相应的随机效应, 揭示了相应成分对不同个体问题解决过程的具体影响程度。结果表明, 较难的项目倾向于单词更多, 命题密度更高, 要求对图/表信息进行编码和转译, 或者根据问题表述生成数学公式。项目命题密度影响效应存在着显著的个别差异。项目命题密度对能力较低的被试的影响高于对能力较高的被试的影响。

关键词 数学应用题; 语言成分; 命题密度; 随机权重 LLTM

分类号 B841

1 引言

1.1 数学应用题的解决过程

数学应用题的理解和解决通常被认为是复杂认知活动, 涉及到多个过程(Mayer, 1985; Schoenfeld, 1983)。基于信息加工理论的视角, Schoenfeld 提出了一个数学问题解决的加工模型。该模型包括五个子过程: 读题(reading)、分析(analysis)、探索(exploration)、计划/执行(planning/implementation)和验证(verification)。承袭同样的研究思路, Mayer 提出了解决数学应用题的四个一般性认知过程, 即问题转译(problem translation)、问题整合(problem integration)、解题计划(solution planning)以及解题执行(solution execution)。问题转译是指把问题陈述转换成数学命题, 例如将文字陈述转换成数学公式。问题整合是指把问题中的各种不同成分, 包括问题陈述、所问以及各种外显或内隐的限制条件等等, 整合成对问题结构的协调表征。数学问题解决计划的制定是建立在对问题形成完整表征基础上的。通过监控并执行该计划, 数学应用题得以解决。当问题是以言语陈述形式呈现时, 在形成对问题中的各种相关信息的整合表征之前, 必须将相关言语陈述正确转化成数学命题。在问题转译阶段, 言语陈述的复杂程度对数学应用题的解决过程起着重要的影响作用。

1.2 数学应用题中的语言成分

研究表明, 对数学问题中言语陈述的理解过程与文本理解的一般原则是一致的(Kintsch & Greeno, 1985)。在已有文本理解的认知研究基础上, Embretson 和 Wetzel (1987)总结了一个阅读理解的认知过程模型, 并将文本理解过程概括为两个主要的子过程, 即语词编码(lexical encoding)和连贯过程(coherence processes)。语词编码意指把文本(符号)从视觉刺激转变成一个有意义的表征的过程, 而连贯过程意指把来自于文本和/或先前知识的各种零散表征整合成连贯的语义网络的过程。语词编码的难度主要受诸如单词数量、单词熟悉度等文本表面结构的影响, 而连贯过程则受文本中命题密度

收稿日期: 2010−09−26

* 上海市浦江人才计划项目资助。

通讯作者: 杨向东, E-mail: xdyang50@hotmail.com

(proposition density)的显著影响。

已有众多研究调查了数学测验项目中言语陈述的表面结构对数学问题解决过程的影响(Abedi & Lord, 2001; Hanson, Hayes, Schriver, leMahieu, & Brown, 1998; Kopriva, 1999; Johnson & Monroe, 2004)。研究表明, 数学题目中语言陈述的某些特定表面特征的改变或简化, 如语义模糊的单词、复杂动词以及数学词汇等, 的确会对项目难度产生一定影响 (Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006)。这种表面特征的变化对语言水平较低的被试有所帮助。然而, 其影响力度通常被认为是很小的(Abedi & Lord, 2001; Johnson & Monroe, 2004)。

命题密度通常指一段文本中所包含的命题数量与字(词)数的比值(Bovair & Kieras, 1985; Embretson & Wetzel, 1987)。已有研究表明, 命题是理解和记忆文本的单位(Kintsch, 1974; Perfetti & Britt, 1995; Turner & Greene, 1977)。命题密度刻画的是文本中所表达的观念(idea)的密集程度, 是阅读者面对文本时所能作出的种种质疑和推断程度的测量指标(Brown, Snodgrass, Kemper, Herman & Covington, 2008)。Kintsch 和 van Dijk (1978)认为, 一个命题由若干概念(concept)构成, 包括一个谓语(predicate), 一个或多个论元(argument)。论元可以是一个概念, 也可以是一个嵌入式命题, 在命题中既可以作主语, 也可以作宾语。谓语通常由动词、修饰语或连接词组成。例如, 句子“John 和 Mary 都喜欢他们的数学老师”可以分解成三个命题(见表 1)。

表 1　例句中的命题

命题序号	命题
P1	(数学, 老师)
P2	(和, John, Mary)
P3	(喜欢, P2,老师)

Embretson 和 Wetzel 区分了三种类型的命题：修饰语命题、谓语命题和连接词命题。例如, 表 1 中的命题 1、2 和 3(表 1 中标示为 P1、P2 和 P3)分别是修饰语命题、连接词命题和谓语命题。其中, 命题 2 作为一个嵌入式命题, 成为命题 3 的一个论元。一般而言, 命题密度与文本理解难度相关。

类似的, Mayer (1985)区分了数学应用题中三种命题类型：赋值(assignment)、关系(relation)和问题(question)。赋值命题给每个变量赋值, 例如“一本书的价格为 65 美元”。关系命题表达了两个变量之间的数量关系, 例如“John 的苹果比 Mary 多 5 个”。问题命题则提出一个关于变量值的问题, 例如“总共是多少？”Mayer 认为, 三种命题类型中, 关系命题对受测者理解题目造成的困扰最大。

2　数学题目中语言成分影响的模型构建

综观已有文献, 存在几种对数学题中语言成分对解题难度的影响的研究模式 (Aiken, 1972; Embretson, 2004; Enright & Sheehan, 2002; Kintsch & Greeno, 1985; Shaftel et al., 2006; Wu & Adams, 2006)。早期研究探讨这个问题, 大多考察被试阅读能力与其解决数学应用题表现之间的关系(详见 Aiken 1972 年的综述文章)。在阅读能力影响数学问题解决的假设下, 早期研究的基本模式如下：(1)测量研究群体(初中学生)的一般阅读能力, (2)测量他们的数学问题解决能力, (3)求两者的相关程度。虽然两者一般呈正相关($0.40 \leqslant r \leqslant 0.86$), 但是如何理解和解释该结果却众说纷纭, 莫衷一是。其中一种主要的观点认为, 被试阅读能力与数学能力之所以存在相关, 是由于两者都与被试的一般智力相关。有研究者建议将一般阅读能力改为特殊阅读能力, 譬如词汇、理解、语法和拼写等(Johnson, 1949)。但是, 这一转变并没有解决根本问题。不同的特殊阅读能力的测量工具或方法, 导致对被试数学问题解决表现的预测效度的不一致。此外, 就理解数学能力与阅读能力之间的关系而言, 单纯依赖两者之间的一阶相关(first-order correlation)或偏相关(partial correlation), 无法提供一个理解特定语言成分是如何影响被试数学问题解决的方法论基础。

另一种方法是通过因素分析研究数学测验题的内在结构, 而不是关注被试阅读能力和数学能力之间的相关模式(DeGuire, 1985; Gorsuch, 1983; Wu & Adams, 2006)。从理论上来说, 如果测试样本在数学题的语言成分上存在实质性的个别差异, 在数学测验的因素结构中应该会找到一个语言因素。反之, 如果样本中所有被试都对数学问题中的语言成分成功的进行了转译, 则在因素结构中不会出现这个语言因素。因此, 无论是探索性因素分析还是验证性因素分析, 都为研究数学题中语言成分影响效应的个体差异提供了一种方法。然而, 从实际数据的因素分析中得出的因素难以解释。某个因素是否真正反应了数学测验中语言成分的影响, 通常是存在争议的。对于同一组数据进行因素分析, 不同的因素抽取方法或者因素旋转方法, 都可能导致不同

的结果, 进而产生不同的解释。即使可以达成一致的解释, 所鉴定出的语言因素仅仅提供了在样本总体变异中, 有多大比例可能是由语言因素导致的。至于到底是语言成分的哪些方面造成了这些个别差异, 它们是如何与数学问题解决的认知过程相联系的(Snow & Lohman, 1989), 因素分析模式下的语言因素是无法回答的。

研究数学应用题中语言成分影响的第三种模式如下：首先, 通过任务分析确认数学问题中言语陈述的具体特征; 然后, 研究这些特征对题目难度以及学生表现的影响效应(Abedi & Lord, 2001; Embretson, 2004; Enright & Sheehan, 2002; Kintsch & Greeno, 1985; Shaftel et al., 2006)。在这种研究模式下有两种不同的研究取向。一种取向聚焦于数学问题中言语陈述的浅层的语言学特征(Abedi & Lord, 2001; Shaftel et al., 2006)。通过改变题目中的非数学词汇和语言结构, 例如词频、动词短语语态等, 研究者希望能在简化语言理解难度的同时, 保留题目中隐含的数学结构。在这种取向下, 言语陈述特征的改变与数学问题解决的认知过程的联系并不清楚。因此, 语言上简化的题目未必就能达成研究者所期望的目的。虽然一般而言, 语言上简化的题目比原题容易, 但是, 在某些情况下, 语言简化的题目甚至出现比原题更难的情况(见 Shaftel 等人 2006 年对相关文献的详细梳理)。

相比之下, 第二种取向下的任务分析充分依托数学问题解决的认知过程模型(Embretson, 2004; Enright & Sheehan, 2002; Kintsch & Greeno, 1985)。例如, 基于 Mayer (1985)的数学问题解决的认知理论, Embretson (2004)分析了 GRE 数学题的心理测量学特征, 旨在鉴别并理解这些题目中涉及的认知复杂水平与来源。运用此种方法时, GRE 数学问题的解决过程被表征为五个不同的阶段, 即编码(encoding)、整合(integration)、计划(planning)、执行(solution execution)和决策(decision)。在此基础上, 鉴别数学问题中对应于每个阶段的认知过程的题目特征集, 并按照 GRE 中的数学问题在题目特征集上的负荷情况对题目进行编码。然后, 运用线性回归模型或线性逻辑斯蒂测验模型(linear logistic test model, LLTM; Fischer, 1973), 对包括言语特征在内的不同题目特征在题目难度上的影响效应进行建模分析和标定。这种基于模型的任务分析方法的一个重要优势在于它是建构导向的(construct-oriented)。通过建立题目特征与问题解决认知过程之间的联系, 并在实证研究的基础上对这些特征的影响效应进行标定, 这种方法提供了研究数学测验背后的建构表征(construct representation)的一种有效途径(Embretson, 1983)。

基于模型的任务分析方法假设不同题目特征对问题解决难度的影响在被试群体中是一个常量。换言之, 题目特征的影响效应被视为固定效应(fixed effect), 暗示着这种效应对所有被试都是相同的。这种假设在某些情况下可能是不现实的。Sebrechts 及其合作者(1996)考察了 GRE 代数应用题的题目特征对题目难度的影响, 并研究了数学问题的陈述方式与被试解决策略或错误之间的关系。结果表明, 不同水平的被试倾向于使用不同的问题解决策略。而且, 在解决数学问题时, 被试是否选择或实施某种具体策略, 似乎部分的取决于该数学问题的某些特定特征, 如问题的语境和语言结构等。错误类型通常与所使用的策略有关。因而, 不同水平的被试产生的错误类型也不同。这表明, 题目特征对被试的影响并不一致, 而是随着题目特征与特定被试的认知特征之间的交互作用而变化。所以, 亟需构想出一种新方法, 以便于分析题目特征对题目难度的这种因人而异的影响效应。

3 本研究的建模方法

本研究中, 数学题目中语言成分的建模分析基于以下的假设：除非样本中所有被试成功渡过问题转译阶段, 数学问题中语言成分的影响会随着被试的不同而不同。换言之, 这种影响在被试群体中被视为随机效应(random effect)。沿袭基于模型的任务分析方法, 本方法始于构建数学问题解决的认知模型, 从中找出与语言成分相关联的认知阶段或过程。然后, 根据相关的研究, 并结合对问题特征的分析, 识别出问题陈述中语言特征, 用以表征与语言成分相关联的认知阶段。此处, 相关研究文献旨在为问题特征选择提供理论基础, 为理解从问题陈述到数学表征的转换过程提供实证依据。

在以上假设基础上, 多元随机效应项目反应模型(IRT models with multiple random effects)可以用于分析语言成分特征在题目难度上的随机效应。目前, 多元随机效应项目反应模型通常被放在广义线性混合模型 (generalized linear mixed model, GLMM) 或非线性混合模型(nonlinear mixed model,

NLMM)的框架下加以理解(Boeck & Wilson, 2004; McCulloch & Searle, 2001; Rijmen, Tuerlinckx, Boeck, & Kuppens, 2003)。在这一框架下, 线性逻辑斯蒂测验模型, 即 LLTM, 可以表述为:

$$\text{logit}\left[P\left(X_{ij}=1|\theta_i,\mathbf{q}_j,\boldsymbol{\eta}\right)\right]=\theta_i-\sum_{k=0}^{K}\eta_k q_{jk} \quad (1)$$

其中, $P\left(X_{ij}=1|\theta_i,\mathbf{q}_j,\boldsymbol{\eta}\right)$ 表示被试 i $(\text{i}=1,2,...\text{I})$ 答对二值计分项目 j $(\text{j}=1,2,...\text{J})$ 的概率。$X_{ij}=1$ 表示正确, 否则 $X_{ij}=0$。θ_i 表示被试 i 的能力参数。$\mathbf{q}_j$ $\left(\mathbf{q}_j=\left\{q_{j0},q_{j1},q_{j2},...,q_{jK}\right\}\right)$ 是所鉴定出的项目特征在项目 j 中的负荷向量。在本研究中, 负荷向量在每个项目中的取值由研究者预先确定。$\boldsymbol{\eta}$ $\left(\boldsymbol{\eta}=\left\{\eta_0,\eta_1,\eta_2,...,\eta_K\right\}\right)$ 向量中包括的是根据实际数据估计出来的项目特征对项目 j 的难度的影响效应, 通常称之为基本参数(basic parameters)。通常设 q_{j0} 为 1, 从而使 η_0 成为截距或标定常数(scaling constant)。公式 1 中, $\text{logit}(\bullet)$ 是 $P\left(X_{ij}=1|\theta_i,\mathbf{q}_j,\boldsymbol{\eta}\right)$ 与 $1-P\left(X_{ij}=1|\theta_i,\mathbf{q}_j,\boldsymbol{\eta}\right)$ 之间比率的对数。

在项目反应理论中,能力参数通常被视为随机效应, 即 θ_i 因人而异。为了解决项目反应模型的界定性问题(identification issue), 通常假设 $\theta_i \sim N\left(0,\sigma_\theta^2\right)$。而项目特征参数通常被视为固定效应。在公式 1 中, $\boldsymbol{\eta}$ 是项目特征参数, 意味着不同项目特征对项目难度的影响效应在被试群体中是不变的。前面提到, 这一假设在某些情况下是不现实。针对这种情况, Rijmen 和 Boeck (2002)提出了随机权重线性逻辑斯蒂测验模型(random-weights linear logistic test model, RW-LLTM)。该模型是 LLTM 的一个变形, 它允许某些或所有基本参数为随机效应, 从而放宽了 LLTM 的假设。该模型可以表述为:

$$\text{logit}\left[P\left(X_{ij}=1|\theta_i,\mathbf{q}_j,\boldsymbol{\eta},\boldsymbol{\lambda}_i\right)\right]=$$
$$\theta_i-\left(\sum_{k=0}^{K_1}\eta_k q_{jk}+\sum_{k=K_1+1}^{K}\lambda_{ik}q_{jk}\right) \quad (2)$$

其中, K_1 是被视为固定效应的项目特征数, K_2 $\left(=K-K_1\right)$ 是被视为随机效应的项目特征数, $\boldsymbol{\lambda}_i$ $\left(\boldsymbol{\lambda}_i=\left\{\lambda_{i(K_1+1)},\lambda_{i(K_1+2)},...,\lambda_{iK}\right\}\right)$ 是随机效应项目特征的基本参数向量。$\boldsymbol{\lambda}_i$ 中下标 i 指 $\boldsymbol{\lambda}_i$ 的值是因人而异的。因此, RW-LLTM 是一个多元随机效应项目反应理论模型 (被试能力 θ_i 也是其中之一)。通常假模型中的随机效应变量符合多元正态分布。

4　方法

4.1　测量工具和数据

测量工具为美国中西部某个州的三年级数学测验。该测验共由 70 道项目, 涵盖了相应课程 4 个主要领域：数与计算、代数、几何和数据分析。分析数据包括 3883 名学生的项目反应。基于所测样本的被试分数计算的测验分数的 α 信度系数为 0.914。

4.2　项目特征及语言成分的识别

数学问题中语言成分的识别遵循了前面所述的基于模型的任务分析模式。本研究以 Mayer (1985)的数学问题解决的认知模型作为任务分析的理论基础。该理论已经得到了广泛地认同与研究(Barnett & Ceci, 2002; Embretson, 1995; Hall, Kibler, & Wenger, 1989; Jonassen, 2003; Lucangeli, Tressoldi, & Cendron, 2002)。如前所述, 该理论将数学应用题解决分解成 4 个认知过程：问题转译、问题整合、解题计划以及解题执行。

参照已有研究的成果(Embretson, 1995, 2004; Embretson and Wetzel, 1987; Sebrechts et al. , 1997; Shaftel et al., 2006), 本研究选择了 6 个项目特征对数学问题的认知过程进行表征, 包括题干字(词)数、命题密度、图表转译、等式生成、子目标数和计算。为了验证项目特征编码的准确性, 两名研究者分别独立的对不同测验项目进行编码, 并就结果进行比较, 所得评分者信度介于 0.88~1.00 之间。表 2 给出了各项目特征的描述及其与数学问题解决的 4 个认知过程的关系。根据理论, 语言成分主要在问题转译阶段对数学问题的解决产生影响。因此, 题干的字(词)数和命题密度被用以表征数学题目中的语言成分。

4.3　建模程序

以测验项目的正确百分率为基础, 求取每个测验项目的 logit 值作为结果变量(outcome variable), 以所选的项目特征为预测变量(predictor), 进行线性回归模型的初步分析。利用回归分析所得的相应变量的回归系数作为比较各种 RW-LLTM 参数估计值的基本参照。这里, 参照的主要目的不是以线性回归分析的结果为评判的最终标准, 而是基于最小二乘法估计的回归系数的无偏性, 并考虑到 RW-LLTM 模型本身及相关估计方法的复杂性, 以回归系数为基础考察相关项目特征影响效应的相对大小和方向。

表 2 基于数学认知过程的项目特征的名称、界定和文献依据

认知阶段/变量	详细描述	文献依据
问题转译		
字(词)数	题干中的单词数量	Embretson & Wetzel (1987) Shaftel et al. (2006)
命题密度	题目中的命题密度	Embretson & Wetzel (1987) Embretson (1995) Kintsch & Greeno (1985) Sebrechts et al. (1996)
问题整合		
图表转译	转译图表(1=是, 0 =否)	Embretson (2004)
等式生成	想出或者转译公式 (1=是,0=否)	Embretson (2004) Daniel & Embretson (2010)
解题计划		
子目标数	题目中目标或子目标的数量	Embretson (2004) Sebrechts et al. (1996) Daniel & Embretson (2010)
解题执行		
计算	所需的数字运算次数	Embretson (2004)

在回归分析的基础上, 对施测数据进行一组基于 RW-LLTM 的随机效应项目反应模型的建模分析。对所有拟合的模型, 假设其中的随机效应变量符合多元正态分布。此外, 出于建模技术上的考虑, 设 $\lambda_{ik}^{*}=\lambda_{ik}-\overline{\lambda}_{k}$, 则有 $\lambda_{ik}^{*}\sim N\left(0,\sigma_{\lambda_{ik}^{*}}^{2}\right)$, 其中 λ_{ik} 是公式 2 中的随机效应参数, $\overline{\lambda}_{k}$ 为对应于 λ_{ik} 的变量 k 的平均固定效应(fixed mean effect)。这样, 每个随机效应的项目特征参数对应两个参数,一个随机效应参数和一个相应的平均固定效应参数。此外, 对每个拟合的 RW-LLTM 模型, 设 $\theta\sim N\left(0,\sigma_{\theta}^{2}\right)$, 其中 σ_{θ}^{2} 是 θ 的随机效应参数。

所拟合的几个 RW-LLTM 模型的依次顺序为: (1)将能力参数 θ_i 作为唯一随机效应的线性逻辑斯蒂测验模型, 即传统的 LLTM; (2)将能力和题干字(词)数作为随机效应的 RW-LLTM; (3)将能力和命题密度作为随机效应的 RW-LLTM; (4)将能力、题干字(词)数和命题密度作为随机效应的 RW-LLTM。所有模型同时控制表 2 中其它的项目特征。所有模型的建模分析均使用 SAS NLMIXED 程序, 采用边缘极大似然估计法(marginal maximum likelihood estimation, MMLE)进行估计(SAS Institute Inc., 2004)。在运用 MMLE 时, 为减少计算时间, 没有选择动态积分点(adaptive quadrature point) 模式。对有一个或两个随机效应的模型, 积分点设定为 10, 对于有三个或等多随机效应的模型, 积分点设定为 5。

上述拟合的一组模型在形式上符合所谓的嵌套模型(nested model)的定义, 即一组模型中简单模型可以通过限定复杂模型中某些参数值而得到(Loehlin, 1998)。嵌套模型拟合程度(goodness-of-fit)的比较通常采用似然比检验(Likelihood-ratio Test)进行(Agresti, 2007)。但是, 似然比检验并不适用于比较具有不同个数的随机效应模型, 即使该组模型符合嵌套的情况(Rijmen & Boeck, 2002)。这是因为随机效应模型的嵌套, 是通过将某个(些)随机效应参数限定在该参数空间的边界(即随机效应方差为零)的基础上实现的。这种情况下的常见做法是利用信息准则(information criterion)选择模型(Schwartz, 1988)。常见信息准则包括 Akaike 信息准则(Akaike information criterion; AIC)或贝叶斯信息准则(Bayesian information criterion; BIC)。已有研究表明, 较之 AIC, BIC 在模型选择上更为一致, 更倾向于选择简化模型(Lin & Dayton, 1997)。因此, 本研究采用 BIC 进行模型选择。一组模型中 BIC 最小的模型即为应选模型(model of choice)。

除了 BIC 之外, 本研究还采用了相对模型拟合指标 Δ^2 (relative model fit index; Embretson, 1997)对模型进行检验。该指标与结构方程建构中的常模化拟合指标 NFI (normed fit index; Bentler & Bonnett, 1980)同理, 表示了当前模型在基准模型(null model)和饱和模型(saturated model)之间的尺度上的相对位置。设 $-2\ln L_n$, $-2\ln L_s$ 和 $-2\ln L_m$ 分别为基准模型、饱和模型和当前模型的–2 log likelihood, 则

$$\Delta^2=\left[\left(-2\ln L_m\right)-\left(-2\ln L_n\right)\right]/\left[\left(-2\ln L_s\right)-\left(-2\ln L_n\right)\right]$$
$$=\left(\ln L_n-\ln L_m\right)/\left(\ln L_n-\ln L_s\right)$$

Δ^2 在原理上类似于线性回归方程中的 R^2, 是当前模型相对拟合程度的数量指标。

5 结果

5.1 项目特征分布情况及线性回归分析结果

整个测验中, 项目题干中平均字(词)数为 20 个左右(M=19.8, SD=11.01), 70%的项目字(词)数少于 22 个。约 65%的项目包含 2~4 个命题, 平均命题密度为 0.18。在 70 个项目中, 近半数项目需要被试对其中的图表进行转译, 或者根据项目文本生成等式。项目所包括的平均子目标数为 2.3 个, 其中 90%的项目包括 2~3 个子目标。绝大多数项目(84.3%)需要的计算少于 3 步。总体而言, 大多数测验项目相对容易(平均正确百分率为 0.83, 最小值和最大值分别为 0.49 和 0.96)。图 1 给出了以正确百分率为指标的项目难度的分布以及经过 logit 转换后的分布情况。可以看出, logit 转换后的项目难度分布更趋向于正态分布, 符合线性回归分析的基本假设。

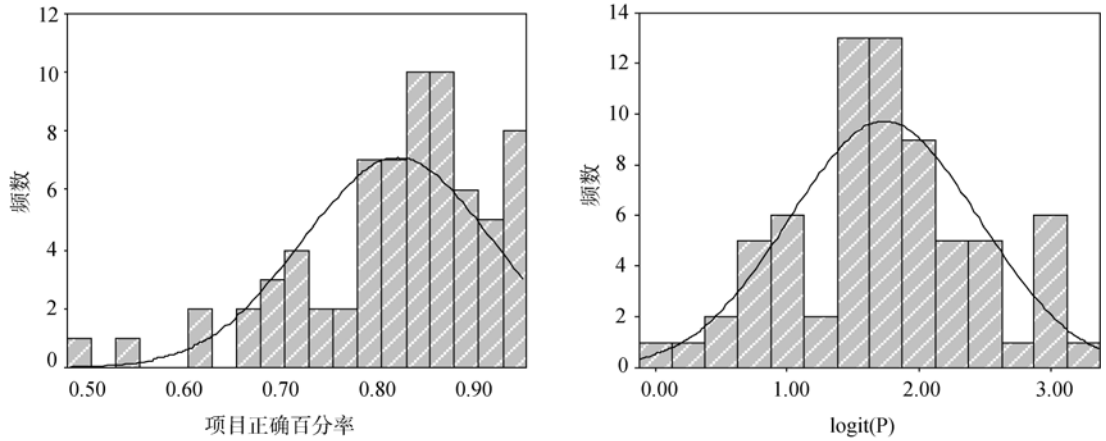

图 1　项目难度的分布情况(上图：以正确百分率为指标的项目难度; 下图：经过 logit 转换后的项目难度)

以经过 logit 转换后的项目正确百分率为结果变量, 表 3 给出了项目特征的参数估计情况。表 3 中的回归系数的解释需要考虑经过 logit 转换后的项目正确百分率的特征。由于容易的项目的正确百分率高, 而 logit 转换并不改变其方向, 因此, 某个项目特征的回归系数为负, 意味着在该变量上负荷高的项目的难度较大。从表 3 可以看出, 除了回归方程的截距和项目特征计算之外, 其他项目特征的回归系数都是负的。这与理论上的预想相一致。难题趋向于具有更多的单词, 命题密度更高, 更多的子目标, 并且可能要求被试对图/表信息进行编码和转译, 或者需要根据问题表述生成数学公式。然而, 除了截距之外, 所有的项目特征对项目难度的影响效应都没有达到统计意义上的显著性。6 个项目特征总共可以解释约 14%的项目难度的变异(R=0.37)。前面提到, 以项目难度为结果变量进行线性回归分析, 旨在为比较各种 RW-LLTM 参数估计值作基本参照。因此, 此处考察的重点是各项目特征的影响效应的大小和方向, 不是影响效应的统计显著性。正如后面的 RW-LLTM 分析所表明的, 随着检验的样本基点不同, 影响效应的统计显著性会随之变化。

表 3　线性回归分析的项目特征的参数估计

项目特征	未经标准化的回归系数		标准化回归系数	t
	估计值	标准误		
截距	2.432	0.327		7.436
字(词)数	–0.103	0.085	–0.158	–1.212
命题密度	–0.990	0.733	–0.167	–1.351
图表转译	–0.222	0.183	–0.156	–1.217
等式生成	–0.310	0.207	–0.215	–1.493
子目标数	–0.019	0.133	–0.019	–0.143
计算	0.002	0.034	0.006	0.048

5.2 基于 RW-LLTM 的建模分析结果

对前面述及的 4 个 RW-LLTM 模型依次利用所测样本的数据进行了估计。同时估计的还有用于计算模型拟合指标 Δ^2 需要的基准模型和饱和模型。表 4 给出了这些模型的参数数量、-2 log likelihood、BIC 以及模型拟合指数 Δ^2。从该表中可以看出, 能力和命题密度作为随机效应的 RW-LLTM 模型(即 RW-LLTM 3)的 BIC 值在所拟合的 4 个模型中最小。该模型的 BIC 值小于将能力、词干和命题密度作为随机效应的模型(即 RW-LLTM 4)的值(相应的 BIC 值分别为 224860 与 225070)。这表明与后者相比, 前者在类似程度拟合数据的基础上, 模型更为简洁。

表 4 不同 RW-LLTM 模型的拟合程度

模型 [a]	模型参数	–2 log likelihood	贝叶斯信息准则 BIC	拟合指数 [b] Δ^2
基准模型	2	226283	226299	--
RWLLTM 1	8	225088	225155	0.059
RWLLTM 2	10	225009	225092	0.063
RWLLTM 3	10	224777	224860	0.075
RWLLTM 4	13	224963	225070	0.065
饱和模型	71	206126	206713	--

注: a. 基准模型是所有项目难度相同的拉希模型(Rasch Model), 饱和模型是传统的拉希模型, RWLLTM 1 是传统的 LLTM 模型, 其中能力为随机效应。RWLLTM 2 是能力和题干字(词)数为随机效应的 RW-LLTM 模型, RWLLTM 3 是能力和命题密度为随机效应的 RW-LLTM, RWLLTM 4 是能力、题干字(词)数和命题密度为随机效应的 RW-LLTM。

b.拟合指数 Δ^2 是当前模型与基准模型的–2 对数似然值之差与当前模型与基准模型的–2 对数似然值之差的比值。

这一选择与模型拟合指数 Δ^2 提供的信息是一致的。上述各种模型都是建立在 LLTM 模型基础上的(其中 RW-LLTM 1 即是以 6 个项目特征参数为基本参数的 LLTM), 而最为简化的 LLTM 模型是只有一个项目截距参数的 LLTM。因此, 基准模型即为只有项目截距参数的 LLTM。该模型假设所有测验项目具有共同的难度系数。另一方面, LLTM 模型是用 K 个项目特征来预测 J 个项目的难度参数($K \leqslant J$) 。当 K=J 时, LLTM 模型即成为传统的拉希模型(Rasch Model)。因此, 选择传统的拉希模型为本研究的饱和模型。根据由此得出的模型拟合指数 Δ^2 来看, 能力和命题密度作为随机效应的 RW-LLTM 模型(即 RW-LLTM 3)的 Δ^2 值最大。这在一定程度上验证了基于 BIC 数值的模型选择。

表 5 能力和命题密度为随机效应的 RW-LLTM 模型的参数估计值

模型	参数		
	估计值 [a]	标准误	t
固定效应			
截距(项目)	2.576	0.029	89.778
字(词)数	–0.098	0.006	–15.856
命题密度	–0.852	0.061	–14.084
图表转译	–0.313	0.012	–25.785
等式生成	–0.368	0.014	–27.032
子目标数	–0.006	0.008	–0.738
计算	0.017	0.002	7.889
随机效应			
	能力 [b]	命题密度	
能力	1.009(0.038)	–0.56(0.068)	
命题密度	**–0.344**	2.630(0.219)	

注: a. 为了便于模型间的对比已将估计值乘以–1.

b. 粗体数字为能力和命题密度之间的相关系数

以能力和命题密度作为随机效应的 RW-LLTM 模型为所选模型, 表 5 给出了该模型中固定效应和随机效应的估计值和标准误差。由于能力参数和项目特征参数之间是对抗性的(即在公式 1 和 2 右侧中的减号), RW-LLTM 模型中参数的方向(即估计值的正负)与前述线性回归的系数正好是相反的。因此, 为了方便与线性回归模型参数的比较, 表 5 中固定效应的估计值都被乘以–1。比较表 3 和表 5 可以看出, 对应项目特征的固定效应的大小和方向是很相似的。进而对项目参数的解释也是与表 3 类似的。然而, 与表 3 不同的是, 基于 RW-LLTM 的估计值除了子目标数之外所有的项目特征对项目难度的影响效应都达到了统计意义上的显著性。这是由于在线性回归分析中样本量的计算是基于项目数的, 而在 RW-LLTM 中是基于被试样本量的, 因而具有更大的统计鉴别力(statistical power)。

另外, 表 5 下半部呈现了所选 RW-LLTM 模型中随机效应的方差-协方差矩阵及其相应的标准误差。能力和命题密度的方差估计值分别为 1.009 和 2.630(根据相应的标准误差, 两个随机效应在统计上都是显著的)。命题密度的效应和被试能力存在一定的负相关, 相关系数为-0.344。这表明, 除了被试能力存在实质性差异之外, 项目中的命题密度在不同被试中也存在着不同的效应。对能力较低的被试而言, 命题密度的作用对被试正确解答项目的影响一般是比较大的。随着能力水平的增加, 这种影响效应逐渐减小。从图 2 中可以明显地看出这一趋势。

5.3 命题密度随机效应的验证

项目命题密度对数学问题解决的影响效应如果随着不同被试而变化, 那么,我们可以做出如下

预测：如果项目命题密度对某些被试的数学问题解决有着重要的影响，那么，对这些被试来讲，项目难度与项目命题密度会有较为密切的关系，项目难度会在一定程度上随着项目命题密度的提高而增加，表现为项目难度与项目命题密度之间有较高的相关。类似的，如果项目命题密度对另外一些被试的数学问题解决的影响不大，则可以预期对这些被试而言，项目难度与项目命题密度的相关关系会减弱。在上述模型中，项目命题密度对被试问题解决的影响是通过个体水平上的命题密度随机效应值来表示的。因此，考察在命题密度随机效应的不同水平上，项目难度与项目命题密度之间的关系，构成了对上述预测的一个检验。

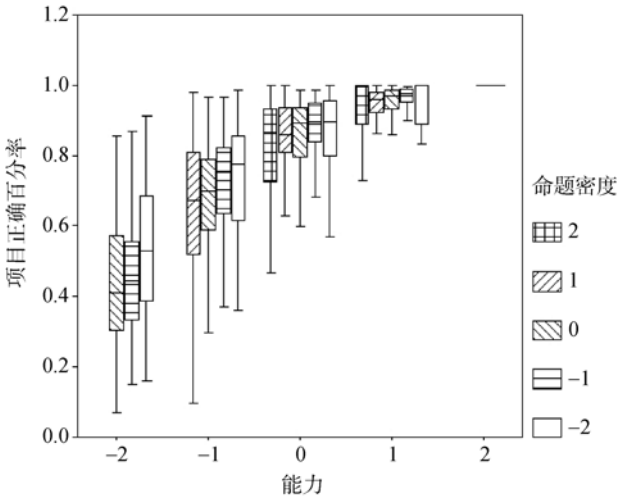

图 2　不同能力水平上命题密度的影响效应

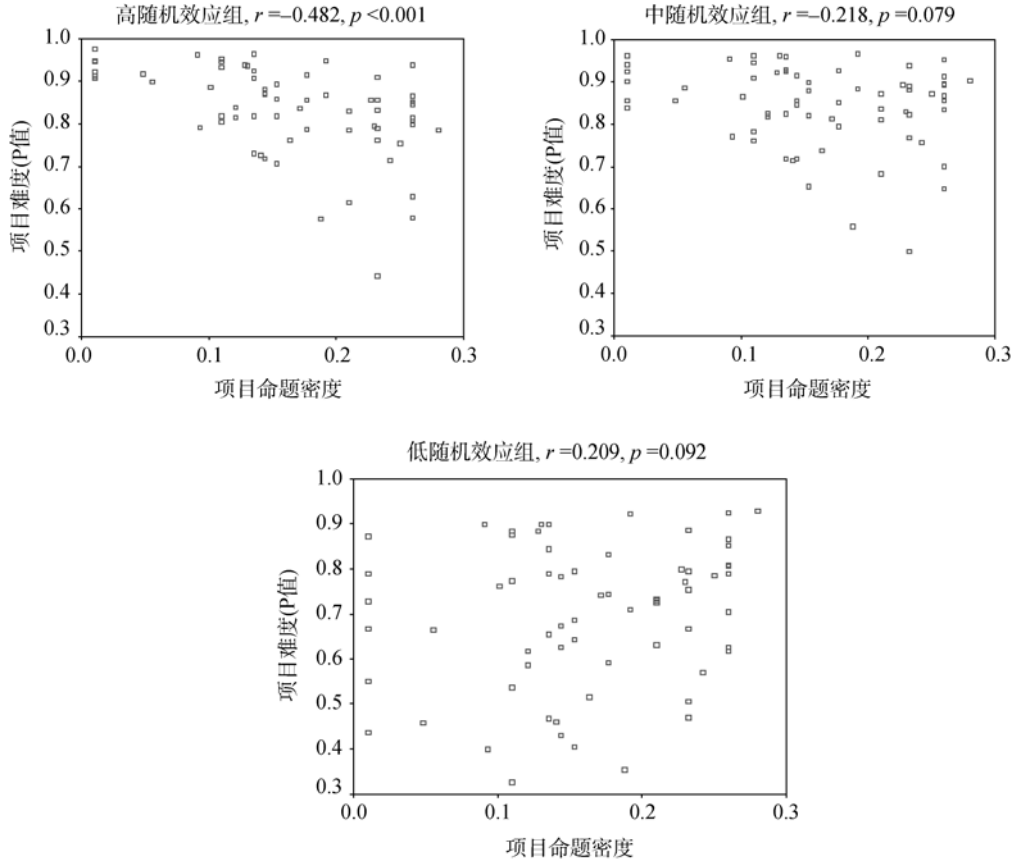

图 3　项目难度(正确百分率)与项目命题密度在不同命题密度随机效应水平上的关系

为了验证这一预测，在所选模型中求取每个被试的命题密度随机效应的经验贝叶斯(empirical Bayes)预测值。然后，按照命题密度随机效应值将被试分为高、中、低三组*。基于不同组的被试的反应数据，考察项目难度与项目命题密度之间的关系。图 3 给出了在不同命题密度随机效应的分组上项目难度与项目命题密度之间的关系。与上述预测相一致，在高命题密度随机效应的被试组，项目难

* 模型中随机效应的分布假设使个体水平上的命题密度随机效应值被重新标定在一个以平均固定效应为均值的正态化尺度上。这种情况下，接近 0 的命题密度随机效应值并不是低随机效应，而是中等随机效应。基于这种思考，高、中、低随机效应的划分标准为：高(随机效应值≥1)，中(–1≤随机效应值≤1)，低(随机效应值≤–1)。

度与项目命题密度之间存在的显著的相关关系(r = –0.482, p<0.001)。对这些被试而言, 问题解决的正确率随着项目中的命题密度的提高而降低。而对其他两组被试而言, 项目难度与项目命题密度之间的相关系数在统计意义上都没有达到显著水平(中随机效应组：r =–0.218, p=0.079; 低随机效应组：r = 0.209, p=0.092), 表明对这两组被试而言, 项目难度与项目命题密度并没有实质意义上的关联。这一结果验证了上述预测的合理性。

6 总结与讨论

本研究分析了数学应用题中语言成分对问题解决的影响。研究在已有认知理论基础上, 鉴别了包括题干字(词)数、命题密度、图表转译、等式生成、子目标数和计算在内的 6 个项目特征对数学问题的认知过程进行表征。其中, 题干字(词)数和命题密度被用以表征数学题目中的语言成分。结果表明, 虽然 6 个项目特征在整体上解释数据变异程度的比重并不大, 但是几乎所有项目特征都对项目难度有着统计意义上的显著效应。较难的项目倾向于单词更多, 命题密度更高, 并且要求对图/表信息进行编码和转译, 或者根据问题表述生成数学公式。此外, 除了被试能力存在实质性差异之外, 项目中的命题密度在不同被试中也存在着不同的效应。命题密度对能力较低的被试的影响高于对能力较高的被试的影响。

本研究与以往研究的不同在于不仅考察了项目中的语言成分对被试数学问题解决的影响的平均效应, 同时分析了这种效应在不同被试之间的个别差异。对该问题的已往研究主要有三种方法：(1)研究受测者的(一般或特殊)阅读能力与数学能力之间的相关关系; (2)通过因素分析来识别一般性的语言因子; (3)通过多元回归模型或线性潜在特质模型, 探讨特定项目特征对项目难度的平均效应。与之相比, 多元随机效应项目反应理论模型提供了一种建构语言成分对数学题目影响的独特方法。与实质理论相结合, 这种方法不仅可以分析影响被试数学问题解决的语言成分的具体特征, 同时还可以分析这些特征与特定被试的认知特性之间的交互作用。这为适当而充分地解释和推断不同被试的测验分数提供了重要信息。

本研究基于数学问题解决过程鉴定出的 6 个项目特征对数据变异的预测程度相对较低。造成这一结果可能有多种原因。首先, 本研究中的项目特征是基于 Mayer (1985)的数学问题解决的认知模型而提出的。虽然该数学问题解决的认知模型已经得到广泛的研究和认可, 但是本研究所用测验项目未必是遵循该认知模型来加以设计的。这种模型与测验编制之间的不匹配有可能导致了相应的项目特征在该测验项目中的低预测力。这里提出了测验编制中的一个重要的理论问题。如果 Mayer 数学问题解决的认知模型代表了认知研究所揭示的数学能力的认知机制, 但与该测验的建构表征并不一致, 那么, 该测验所测量的数学能力又是什么？其测量建构中的表征不足(construct under-representation)和无关变异(construct-irrelevant variance)的因素又有哪些(Messick, 1995)？后继研究可以通过对具体测验项目的结构或认知分析, 对这些问题做出进一步的回答。其次, 本研究对数学能力的建构表征的检验, 是以随机效应项目反应理论模型为纽带, 通过选择项目任务特征来实现对测验建构的认知理论的操作化表征, 从而达到检验的目的的。虽然本研究在项目特征的选择上以相关研究为依据, 并对项目编码的准确性做了检验, 但中间环节的失真仍有可能造成对项目难度变异预测程度的降低。第三, 本研究所要测验项目相对被试群体而言比较简单, 从而导致项目特征与项目难度之间的关系降低。这在一定程度上也会影响项目特征的预测能力。后继研究需要在控制测验项目难度与被试群体能力水平的对应程度的基础上深化对该问题的研究。

参 考 文 献

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219–234.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd edition. John Wiley & Sons.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Educational Research, 42*, 359–385.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637.

Bentler, P. M., & Bonnett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588 –606.

Boeck, P. D., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Bovair, S., & Kieras, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (eds.). *Understanding expository text*. (pp. 315–362). Hillsdale, NJ: Erlbaum.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging.

Behavior Research Methods, 40, 540–545.

DeGuire, L. J. (1985). *The structure of mathematical abilities: the view from factor analysis*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Embretson, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: application to mathematical reasoning. *Journal of Educational Measurement, 32*, 277–294.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden. & R. K. Hambleton. (Eds.), *Handbook of modern item response theory*. (pp. 305–322). New York, NY: Springer.

Embretson, S. E. (2004). *A cognitive model of mathematical reasoning*. Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology. Naples, FL: October.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement, 11*, 175–193.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: implications for item generation. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development,* (pp. 129 –158). Mahwah, NJ: Lawrence Erlbaum Associates.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. *Cognition and Instruction, 6*, 223–283.

Hanson, M. R., Hayes, J. R., Schriver, K., LeMahieu, P. G., & Brown, P. J. (1998). *A plain language approach to the revision of test items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Johnson, J. T. (1949). On the nature of problem solving in arithmetic. *Journal of Educational research, 43*, 110–115.

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29*, 35–45.

Jonassen, D. H. (2003). Designing research-based instruction for story problems. *Educational Psychology Review, 15*, 267–296.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*, 109–129.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Kopriva, R.J (1999). *Ensuring accuracy in testing for English language learners: A practical guide for assessment development*. Washington, DC: Council of Chief State School Officers.

Lin, T. H., & Dayton, C. M. (1997). Model-selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*, 249–264.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path and structural analysis*, 3rd edition. Hillsdale, NJ: Erlbaum.

Lucangeli, D., Tressoldi, P., & Cendron, M. (2002). Cognitive and metacognitive abilities involved in the solution of mathematical word problems: validation of a comprehensive model. *Contemporary Educational Psychology, 23*, 257–275.

Mayer, R. E. (1985). Mathematical ability. In R. J. Sternberg. (Ed.), *Human abilities: Information processing approach* (pp. 127–150). San Francisco, CA: Freeman.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50* (9), 741–749.

Perfetti, C. A., & Britt, M. A. (1995). Where do propositions come from? In C. A. Weaver, S. Mannes, & C. R. Fletcher. (eds.), *Discourse comprehension: essays in honor of Walter Kintsch*. (pp. 11–34). Hillsdale, NJ: Erlbaum.

Rijmen, F., & Boeck, P. D. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26,* 271–285.

Rijmen, F., Tuerlinckx, F., Boeck, P. D., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185–205.

SAS Institute Inc. (2004). *SAS online doc 9.13*. Cary, NC: SAS Institute Inc.

Schoenfeld, A. H. (1983). Episodes and executive decisions in mathematical problem solving. In R. Lesh & M. Landau, M. (Eds.). *Acquisition of mathematics concepts and processes* (pp. 345–395). New York, NY: Academic.

Sebrechts, M. M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using Algebra word problem to assess quantitative ability: Attributes, Strategies, and Errors. *Cognition and Instruction, 14*, 285 –343.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and student with disabilities. *Educational Assessment, 11*, 105–126.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement*, 3rd edition, (pp. 263–331). New York, NY: American Council on Education /Macmillan.

Turner, A., & Greene, E. (1977). *The construction and use of a propositional text base* (Technical Report 63). University of Colorado, Institute for the Study of Intellectual Behavior; Boulder, CO.

Wu, M., & Adams, R. (2006). Modeling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18*, 93–113.

Modeling Language Components in Mathematical Items Using Multiple Random Effects IRT Models

YANG Xiang-Dong[1]; CHIU Lorraine[2]; WEI Yun[1]

([1] *Institute of Curriculum and Instruction, East China Normal University, Shanghai* 200062, *China*)

([2] *Department of Psychological Research in Education, University of Kansas, USA*)

Abstract

Language components in mathematical word problems may have profound impacts on the problem-solving processes engaged by examinees. When a math problem is presented by means of language statements, translation of these language statements into mathematical propositions must be carried out properly before relevant information can be integrated into a coherent representation of the problem. As many studies have shown, the impacts of such language components are unlikely to occur uniformly across examinees. Yet existing approaches that address this issue focus only on (1) examining the first-order correlation between the reading and mathematical abilities of examinees, (2) identifying a general language factor through factor analysis, and (3) examining the average impacts of specific language features on math item properties through linear regression or linear logistic test model (LLTM). While the first two approaches fall short of providing targeted information about the specific aspects of language components that impact examinees' mathematical problem solving, the third approach is based on an unrealistic assumption of constant impacts of such language components. Therefore, an alternative approach needs to be formulated in order to model individual-specific impacts of particular language features on math item proficiency.

The current study applies item response theory (IRT) models with multiple random effects to investigate the effects of language components in mathematical items on examinees' performances. This approach starts by studying relevant literature and searching for a cognitive processing model for mathematical problem solving, and then identifies specific item stimulus features in the problem statements to represent the language components according to the model. By encoding items in a third-grade mathematical test according to the set of identified stimulus features, a set of models that operationalize different cognitive principles are then applied to the data obtained from the test. This approach provides a rigorous method for examining not only the average impacts of specific language features on item properties over the sample, but also variations of such impacts among examinees. Consequently, it allows researchers to investigate the interactions between the characteristics of a mathematical problem and the cognitive abilities of a particular examinee.

Results from this study show that, among the six item stimulus features that were identified based on Mayer's cognitive theory of mathematical problem solving and its associated literature, all but one significantly affect the difficulty of mathematical items. Difficult items tend to have more words, feature a higher proposition density, and require the examinee to either translate information from a graph or table or to generate mathematical equations from the problem statements. In addition to the variation of mathematical ability across examinees, proposition density of the mathematical items shows differentiated impacts for different examinees in the testing sample. While the random effects of proposition density inversely relate to the mathematical abilities of examinees in general, proposition density exerts more impacts on the math problem solving for examinees with low abilities, and such impact is diminished as the examinee's ability increases.

Key words math word problem; language component; proposition density; RW-LLTM