认知诊断模型中题目拟合评估的研究*

高旭亮** 王 芳 夏林坡 侯敏敏
（贵州师范大学心理学院，贵州师范大学心理健康教育与咨询中心，贵阳，550025）

摘 要 有效应用认知诊断模型（cognitive diagnosis model, CDM）的一个关键步骤是检查模型和测验题目是否拟合。尽管已有研究将 IRT 中的题目拟合检验方法应用于 CDM 中，然而这些方法在 CDM 中的表现仍缺乏系统的比较研究。本研究通过模拟实验比较了 χ^2，G^2，$S-\chi^2$，$z(r)$，$z(l)$ 和 Stone–Q_1 的一类错误率和统计检验力。实验结果显示，综合一类错误率和统计检验力而言，当用 ACDM 作为生成模型时，$z(r)$ 和 $z(l)$ 的效果最优；当生成模型是 DINA 或 DINO 时，在高质量测验中，$z(r)$ 的表现最好，而在低质量测验中，χ^2 和 G^2 的表现更好。最后通过一个实测数据分析，进一步检验了题目拟合检验方法的实证应用效果。

关键词 认知诊断模型 题目拟合 一类错误率 统计检验力

1 引言

认知诊断模型（cognitive diagnosis model, CDM）可以提供关于每个学生在学习相关属性方面的优势和劣势的重要诊断信息（Li et al., 2020）。尽管 CDM 最初是被应用于教育评估领域，但它现在正被用于评估其他类型的结构，如心理障碍（de la Torre et al., 2018; Xi et al., 2020）和基于情境的能力评估（Sorrel et al., 2016）。当前，基于对解决问题过程的不同认知假设，即认知过程、技能或属性如何影响学生对项目的作答反应，已有学者开发了不同类型的 CDM。像任何基于模型的评估一样，有效应用CDM的关键一步是检查模型和数据的拟合度，即模型预测与观察数据之间的一致性（Wang et al., 2015）。当模型与数据不拟合时，使用模型估计的参数进行推断的有效性会受到很大的影响。

评估模型和数据的拟合，通常需要同时评估测验水平拟合（test-level fit）和题目水平拟合（item-level fit）两个方面。测验拟合从总体水平上评估模型和数据的拟合，测验拟合通常是比较几个模型在同一批数据的相对拟合（relative fit）统计量。题目拟合用于评估每个题目和模型的拟合度，有助于识别异常题目，通过删除或修改异常题目将提高整个测验和模型的拟合水平（Wang et al., 2015）。换句话说，题目水平的拟合分析不仅是对测验水平拟合的补充检验，而且在心理和教育测量工具开发中也是必不可少的，因为题目拟合结果将有助于指导题目的修订或删除（Liu & Maydeu-Olivares, 2014）。

在 IRT 框架下，已有大量关于题目拟合检验的研究（Chalmers & Ng, 2017; Köhler et al., 2020; Su et

* 本研究得到贵州省科技计划项目（黔科合基础 –ZK[2021] 一般 123）、贵州省高校人文社会科学研究项目（2020QN018）和贵州师范大学 2019 年博士科研启动项目（GZNUD[2019] 27 号）的资助。

** 通讯作者：高旭亮，E–mail: gaoxl9817@foxmail.com

DOI:10.16719/j.cnki.1671–6981.20240226

al., 2021; Zhang et al., 2018）。但是，在认知诊断理论下，关于题目拟合检验的研究仍然不多。当前，仅有少量研究试着将 IRT 中的题目拟合检验指标拓展到 CDM 中，例如，涂冬波等人（2014）比较了 χ^2 和 G^2 统计量在 DINA 模型的效果；Wang 等人（2015）将 IRT 中的题目拟合指标：Q_1 和 PD（power-divergence）等应用于 DINA 模型中；Sorrel 等人（2017）将 S-χ^2 应用于 CDM 中；Chen 等人（2013）将基于题目对（item pairs）的统计量应用于 CDM 中。然而，一方面，已有的研究主要集中在 DINA 模型下，比较传统题目拟合方法的效果，而这些题目拟合方法在其他 CDM 下的效果如何，仍值得探讨；另一方面，上述这些题目拟合方法都属于绝对题目拟合（absolute item fit）指标，绝对题目拟合在实际应用中也是最常用的一类模型拟合评价方法，例如在 IRT 的应用中，有大量的研究使用 S-χ^2 指标来评估题目拟合（Acevedo-Mesa et al., 2020; Flens et al., 2019）。尽管这些绝对题目拟合方法已被初步应用于 CDM 中，但这些方法在 CDM 的效果仍缺乏系统比较，在 CDM 的题目拟合检验中，这些指标的效果如何？面对不同的测验情境，该如何选择最佳的题目拟合检验指标？因此，本研究旨在不同的实验条件下，系统比较这些绝对题目拟合方法在 CDM 的表现，从而为实际使用者在题目拟合方法的选用上提供有价值的参考。

2 题目拟合检验指标

2.1 χ^2 和 G^2 统计量

在 CDM 题目拟合检验中，χ^2 和 G^2 的基本思路是按照估计的被试属性模式将其分类，并通过比较分类数据的观测频率和期望频率来计算题目拟合，χ^2 和 G^2 的计算公式分别如下：

$$\chi_j^2=\sum_{l=1}^{2^K}N_l\frac{\left(O_{jl}-E_{jl}\right)^2}{E_{jl}(1-E_{jl})} \quad (1)$$

$$G_j^2=2\sum_{l=1}^{2^K}N_k\left[O_{jl}\ln\left(\frac{O_{jl}}{E_{jl}}\right)+\left(1-O_{jl}\right)\ln\left(\frac{1-O_{jl}}{1-E_{jl}}\right)\right] \quad (2)$$

公式（1）和（2）中的 K 表示考察的属性个数，$\boldsymbol{\alpha}_l$ 表示被试的属性模式，它共包括 2^K 种类别。N_l 表示属性模式为 $\boldsymbol{\alpha}_l$ 的被试人数，O_{jl} 表示属性模式为 $\boldsymbol{\alpha}_l$ 的被试在第 j 题观察答对的人数，E_{jl} 表示属性模式为 $\boldsymbol{\alpha}_l$ 的被试期望答对第 j 题的人数。χ^2 和 G^2 近似服从自由度为 2^K-m 的卡方分布，m 表示使用 CDM 估计的题目参数个数。

2.2 $S-\chi^2$ 统计量

S-χ^2 统计量按照测验得分对被试进行分类，通过比较每个累加分数的观察频率和期望频率的差异来评估题目拟合，S-χ^2 的计算如下：

$$S-\chi_j^2=\sum_{s=1}^{J-1}N_l\frac{\left(O_{js}-E_{js}\right)^2}{E_{js}(1-E_{js})} \quad (3)$$

其中 s 表示以测验总分划分的分组，J 表示题目个数，O_{js} 和 E_{js} 分别表示第 s 组被试在题目 j 上实际答对的比例和期望答对的比例。E_{js} 的定义如下：

$$E_{jk}=\frac{\sum_{l=1}^{2^K}P(x_{ij}=1|\boldsymbol{\alpha}_l)P(S_i^j=s-1|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K}P(S_i=s|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)} \quad (4)$$

上式中，$P(S_i^j=s-1|\boldsymbol{\alpha}_l)$ 表示被试已作答 j-1 个题目后，恰得 s-1 分的概率。它可以用递归算法计算得到（Orlando & Thissen, 2000），$p(\boldsymbol{\alpha}_l)$ 表示属性模式 $\boldsymbol{\alpha}_l$ 的分布概率。S-χ^2 统计量近似服从自由度为 J-m-1 的卡方分布，m 是题目参数的个数。

2.3 Stone–Q_1 统计量

Q_1 统计量的计算公式如下：

$$Q_1=\sum_{l=1}^{2^K}N_l\frac{\left(O_{jl}-E_{jl}\right)^2}{E_{jl}(N_l-E_{jl})} \quad (5)$$

上式中 N_l，O_{js} 和 E_{js} 分别表示属性模式为 $\boldsymbol{\alpha}_l$ 的总人数以及在第 j 题上观察答对和期望答对的人数。Stone-Q_1 统计量是 Q_1 和 Monte Carlo 重复采样技术的结合，具体过程为：根据真实作答数据估计的题目参数，重新模拟作答数据，基于模拟作答数据计算 Q_1。重复这个过程多次，用于构建 Q_1 统计量的抽样分布，从而检验题目是否拟合。

2.4 基于题目对的统计量

Chen 等人（2013）提出了基于题目对的拟合检验量：转换相关性 $r_{jj'}$ 和对数比值比 $l_{jj'}$。令 $\mathbf{X}_j=\{X_{1j},...,X_{ij},...,X_{Nj}\}$ 表示 N 个真实被试在题目 j 的作答向量，从估计得到的属性模式后验分布中进行取样，可以模拟生成 $\mathring{N}$ 个被试的属性模式，$\mathring{\mathbf{X}}_j=\{\mathring{X}_{1j},...,\mathring{X}_{ij},...,\mathring{X}_{\mathring{N}j}\}$ 表示 $\mathring{N}$ 个被试在题目 j 的模拟作答向量。基于以上定义，$r_{jj'}$ 的计算公式可以写作：

$$r_{jj'}=\left|Z[Corr(\mathbf{X}_j,\mathbf{X}_{j'})]-Z[Corr(\mathring{\mathbf{X}}_j,\mathring{\mathbf{X}}_{j'})]\right| \quad (6)$$

这里 $corr(\cdot)$ 表示题目对之间的皮尔逊相关，$Z[\cdot]$ 表示 Fisher-z 转换，$r_{jj'}$ 的标准误计算公式定义如下：

$$SE[r_{jj'}]=[N-3]^{1/2} \quad (7)$$

对数比值比$l_{jj'}$的计算公式如下：

$$l_{jj'}=\left|\log\left(\frac{N_{11}N_{00}}{N_{01}N_{10}}\right)-\log\left(\frac{\overset{\circ}{N}_{11}\overset{\circ}{N}_{00}}{\overset{\circ}{N}_{01}\overset{\circ}{N}_{10}}\right)\right| \quad (8)$$

$N_{yy'}$表示观察到的在题目 j 和 j' 上得分为 y 和 y' 分的人数，$\overset{\circ}{N}_{yy'}$表示期望在题目 j 和 j' 上得分为 y 和 y' 分的人数，其中，$j \neq j'$且$y, y' \in \{0,1\}$。$l_{jj'}$指标的标准误计算公式如下：

$$SE[l_{jj'}]=[\overset{\circ}{N}(1/\overset{\circ}{N}_{11}+1/\overset{\circ}{N}_{00}+1/\overset{\circ}{N}_{10}+1/\overset{\circ}{N}_{01})/N]^{1/2} \quad (9)$$

已知标准误，可以分别计算$r_{jj'}$和$l_{jj'}$统计量对应的 Z 分数。为了减少计算量，Chen 等人（2013）提出选取每个题目中相应统计量的最大 z 分数，通过最大 z 分数对应的 p 值来评价题目是否拟合。同时，为了降低多次比较引起的一类错误率膨胀，Chen 等人（2013）建议对 p 值进行校正，本研究使用 Holm 调整法（Holm' s adjustment）对 p 值加以校正。为了方便描述，本文将转换之后的统计量简记为 $z(r)$ 和 $z(l)$，文章接下来的部分都用 $z(r)$ 和 $z(l)$ 来代替经过 z 分数转化后的相关统计量。

3 蒙特卡洛模拟实验

3.1 实验设计

为了系统地调查题目拟合统计指标在不同 CDM 的效果，设计了一个蒙特卡洛模拟实验。实验自变量包括：（1）样本容量设置了 2 种：N=500, 1000，被试的属性模式从高阶分布（de la Torre & Douglas, 2004）中生成；（2）生成模型，包括 DINA，DINO 和 ACDM；（3）显著性水平，包括 2 种水平 α = .01 和 α = .05；（4）拟合模型：DINA，DINO 和 ACDM；（5）题目拟合检验指标：χ^2，G^2，S-χ^2，$z(r)$，$z(l)$ 和 Stone-Q_1；（6）测验长度，包括 30 和 60 题，60 题由重复 30 题 2 次而生成。参考 Ma 等人（2016）的模拟方法，测验考察了 5 个属性，包含 30 个题目，测验的 Q 矩阵见表 1；（7）测验质量（高质量和低质量）。题目参数的模拟参考 Gao 等人（2020）的方法，用 $P(1)$ 和 $P(0)$ 来代替模拟 3 个生成 CDM 的原始参数，即对于高质量和低质量的题目，1-$P(1)$ 和 $P(0)$ 分别从均匀分布 U（ .05, .15）和 U（ .15, .25）随机生成；$P(1)$ 和 $P(0)$ 分别表示被试在题目上的最高和最低答对概率。

采用 3 种生成 CDM 模拟被试作答，然后再次使用 3 种生成 CDM 拟合数据，例如采用 DINA 模型生成数据，分别使用 DINA、DINO 和 ACDM 拟合数据。每种实验条件重复模拟 100 次，因变量为一类错误率和统计检验力。为了便于读者理解，现以一个生成模型为例，说明题目拟合指标的一类错误率和统计检验力的计算过程。一类错误率是指当生成模型和拟合模型一致的情况下，统计方法拒绝生成模型（拟合模型）的比例。例如，当生成模型是 DINA 时，一类错误率是指在 100 次重复实验中，每个题目拟合方法拒绝 DINA 的比例；而统计检验力是指生成模型和拟合模型不一致时，统计方法拒绝拟合模型的比例，如当生成模型是 DINA 时，使用 DINO 和 ACDM 模型拟合数据，DINO 模型对应的统计检验力是指成功拒绝 DINO 的比例。

3.2 实验结果

3.2.1 一类错误率

表 2 和 3 分别汇总了低质量和高质量条件下，6 种题目拟合指标的一类错误率，表中第一个平均值

表 1 模拟测验的 Q 矩阵

题目	A1	A2	A3	A4	A5	题目	A1	A2	A3	A4	A5
1	1	0	0	0	0	16	1	1	1	0	0
2	0	1	0	0	0	17	1	1	0	1	0
3	0	0	1	0	0	18	1	1	0	0	1
4	0	0	0	1	0	19	1	0	1	1	0
5	0	0	0	0	1	20	1	0	1	0	1
6	1	1	0	0	0	21	1	0	0	1	1
7	1	0	1	0	0	22	0	1	1	1	0
8	1	0	0	1	0	23	0	1	1	0	1
9	1	0	0	0	1	24	0	1	0	1	1
10	0	1	1	0	0	25	0	0	1	1	1
11	0	1	0	1	0	26	1	1	1	1	0
12	0	1	0	0	1	27	1	1	1	0	1
13	0	0	1	1	0	28	1	1	0	1	1
14	0	0	1	0	1	29	1	0	1	1	1
15	0	0	0	1	1	30	0	1	1	1	1

表 2 低质量测验下题目拟合方法的一类错误率

L	N	统计量	α = .01 DINA	DINO	ACDM	平均值	α =.05 DINA	DINO	ACDM	平均值
30	500	$z(r)$	.006	.008	.002	.005	.026	.029	.006	.020
		$z(l)$	.006	.007	.002	.005	.026	.026	.008	.020
		Stone-Q_1	.014	.009	.013	.012	.063	.058	.058	.060
		S-χ^2	.023	.021	.034	.026	.068	.059	.096	.074
		χ^2	.036	.033	.011	.027	.097	.093	.059	.083
		G^2	.007	.006	.006	.006	.039	.031	.044	.038
	1000	$z(r)$	.005	.003	.002	.003	.028	.026	.010	.021
		$z(l)$	.004	.004	.002	.003	.026	.016	.010	.018
		Stone-Q_1	.015	.014	.012	.014	.066	.064	.064	.065
		S-χ^2	.024	.029	.036	.030	.069	.080	.106	.085
		χ^2	.081	.082	.013	.059	.152	.150	.081	.127
		G^2	.015	.018	.011	.015	.065	.066	.074	.068
60	500	$z(r)$	.006	.005	.003	.005	.031	.027	.020	.026
		$z(l)$	.007	.004	.003	.004	.024	.026	.021	.024
		Stone-Q_1	.010	.011	.009	.010	.042	.042	.042	.042
		S-χ^2	.008	.009	.012	.010	.039	.039	.059	.046
		χ^2	.015	.013	.010	.013	.052	.056	.047	.052
		G^2	.006	.005	.006	.006	.033	.030	.027	.030
	1000	$z(r)$	.006	.003	.003	.004	.030	.025	.014	.023
		$z(l)$	.005	.001	.003	.003	.023	.019	.013	.018
		Stone-Q_1	.008	.011	.011	.010	.045	.044	.046	.045
		S-χ^2	.008	.010	.018	.012	.034	.045	.069	.050
		χ^2	.009	.010	.008	.009	.049	.051	.048	.049
		G^2	.003	.005	.006	.005	.031	.034	.030	.032

表示在 α = .01 条件下，3 种模型的平均一类错误率，第二个平均值表示在 α = .05 条件下，3 种模型的平均一类错误率。

由表 2 的结果可知，当显著性水平为 .01 时，在 30 题的条件下，Stone-Q_1，S-χ^2，$z(r)$ 和 $z(l)$ 的一类错误率在不同样本容量中的表现稳定，其中 $z(r)$ 和 $z(l)$ 的平均一类错误率最低，始终保持在 .01 以下，Stone-Q_1 的平均一类错误率略低于 S-χ^2。当测验长度增加到 60 题时，Stone-Q_1，S-χ^2，$z(r)$ 和 $z(l)$ 的一类错误率并没有明显的变化。但随着题目长度的增加，χ^2 和 G^2 的平均一类错误率有下降的趋势，显然增加测验长度有利于降低和 G^2 所犯的一类错误率。

当显著性水平为 .05 时，在不同测验长度下，$z(r)$ 和 $z(l)$ 的平均一类错误率始终最低，并且维持在 .04 以下；而其他拟合方法，在同等样本容量下，随着测验长度的增加，一类错误率也随之降低。例如，当 N=1000 时，在测验长度为 30 的条件下，Stone-Q_1，S-χ^2，和 G^2 的平均一类错误率分别是 .065，.085，.127 和 .068，当测验长度为 60 题时，平均一类错误率则降低为 .045，.050，.049 和 .032。这表明 $z(r)$ 和 $z(l)$ 的所犯的一类错误率最低，并且不会随着题目数量的变化而有较大的波动。

由表 3 的结果可发现，当测验长度变化时，在同等的样本容量下，这 6 种题目拟合指标的表现保持稳定，$z(r)$ 和 $z(l)$ 的平均一类错误率始终是最低的，Stone-Q_1 和 G^2 的平均一类错误维持在 .01 左右，S-χ^2 的平均一类错误率要略高于 Stone-Q_1 和 G^2，而 $z(r)$ 和 $z(l)$ 的平均一类错误率稳定在 .03 以下。

而当显著性水平为 .05 时，相同条件下，每个题目拟合指标的一类错误率也会增加，$z(r)$ 和 $z(l)$ 的平均一类错误率始终是最低的，其中，$z(l)$ 的一类错误率要略低于 $z(r)$。S-χ^2 和 χ^2 的平均一类错误率都超过了 .05，Stone-Q_1 的平均一类错误率始终小于 .05。当 L=60 时，不同样本容量下，题目拟合方法的平均一类错误率几乎保持一致。

3.2.2 统计检验力

表 4 和 5 分别汇总了生成模型为 ACDM 时，6 种题目拟合指标在低质量和高质量测验条件下的统计检验力。表 6 和 7 分别汇总了生成模型为 DINA 时，在低质量和高质量测验条件下的统计检验力。表 8 和 9 分别汇总了生成模型为 DINO 时，在低质量和高质量测验条件下的统计检验力。下列表中两个平均值分别表示在 α = .01 和 .05 的条件下，题目拟合

表 3 高质量测验下题目拟合方法的一类错误率

			α=.01			平均值	α=.05			平均值
L	*N*	统计量	DINA	DINO	ACDM		DINA	DINO	ACDM	
30	500	$z(r)$	.004	.004	.002	.003	.014	.014	.006	.011
		z(l)	.003	.004	.002	.003	.013	.010	.008	.010
		Stone-Q_1	.011	.009	.011	.010	.050	.037	.057	.048
		S - χ^2	.028	.030	.040	.033	.069	.068	.111	.083
		χ^2	.024	.016	.016	.019	.080	.068	.062	.070
		G^2	.005	.006	.005	.005	.030	.022	.035	.029
	1000	$z(r)$	.002	.003	.001	.002	.007	.013	.009	.010
		z(l)	.001	.002	.001	.001	.008	.008	.009	.008
		Stone-Q_1	.008	.010	.013	.010	.043	.048	.054	.048
		S - χ^2	.016	.016	.032	.021	.051	.054	.094	.067
		χ^2	.034	.036	.012	.027	.089	.098	.060	.082
		G^2	.005	.004	.006	.005	.028	.028	.036	.030
60	500	$z(r)$	.003	.003	.000	.002	.012	.012	.006	.010
		z(l)	.002	.003	.000	.002	.012	.010	.007	.010
		Stone-Q_1	.009	.010	.008	.009	.046	.045	.050	.047
		S - χ^2	.029	.031	.029	.030	.064	.066	.082	.071
		χ^2	.013	.014	.013	.013	.055	.055	.059	.056
		G^2	.003	.003	.005	.004	.025	.025	.032	.027
	1000	$z(r)$	.003	.002	.001	.002	.013	.017	.007	.012
		z(l)	.001	.002	.001	.001	.006	.011	.007	.008
		Stone-Q_1	.010	.008	.009	.009	.043	.044	.051	.046
		S - χ^2	.028	.028	.037	.031	.066	.066	.094	.075
		χ^2	.011	.012	.010	.011	.055	.054	.051	.053
		G^2	.003	.003	.005	.004	.025	.025	.030	.026

表 4 低质量测验下生成模型为 ACDM 的统计检验力

			α=.01		平均值	α=.05		平均值
L	*N*	统计量	DINA	DINO		DINA	DINO	
30	500	$z(r)$	.350	.379	.364	.509	.545	.527
		z(l)	.336	.360	.348	.499	.529	.514
		Stone-Q_1	.148	.174	.161	.286	.327	.306
		S - χ^2	.038	.056	.047	.113	.134	.123
		χ^2	.169	.185	.177	.306	.333	.320
		G^2	.145	.152	.149	.276	.279	.277
	1000	$z(r)$	.634	.676	.655	.757	.810	.784
		z(l)	.626	.665	.645	.748	.804	.776
		Stone-Q_1	.289	.315	.302	.450	.460	.455
		S - χ^2	.070	.117	.094	.174	.249	.212
		χ^2	.383	.386	.384	.559	.535	.547
		G^2	.346	.348	.347	.508	.494	.501
60	500	$z(r)$	.529	.530	.529	.676	.698	.687
		z(l)	.519	.509	.514	.665	.683	.674
		Stone-Q_1	.310	.299	.305	.475	.449	.462
		S - χ^2	.042	.051	.046	.126	.155	.141
		χ^2	.331	.313	.322	.481	.453	.467
		G^2	.303	.277	.290	.437	.402	.420
	1000	$z(r)$	.778	.827	.803	.865	.917	.891
		z(l)	.768	.818	.793	.857	.912	.884
		Stone-Q_1	.540	.486	.513	.679	.606	.643
		S - χ^2	.106	.142	.124	.243	.296	.270
		χ^2	.578	.506	.542	.700	.626	.663
		G^2	.554	.494	.524	.676	.604	.640

表 5 高质量测验下生成模型为 ACDM 的统计检验力

			α=.01		平均值	α=.05		平均值
L	*N*	统计量	DINA	DINO		DINA	DINO	
30	500	*z(r)*	.855	.900	.878	.914	.945	.930
		z(*l*)	.848	.890	.869	.905	.942	.924
		Stone-Q_1	.496	.466	.481	.637	.571	.604
		$S-\chi^2$	.214	.241	.228	.376	.381	.378
		χ^2	.556	.500	.528	.688	.611	.650
		G^2	.520	.464	.492	.642	.563	.603
	1000	*z(r)*	.948	.981	.965	.974	.992	.983
		z(*l*)	.934	.979	.957	.960	.990	.975
		Stone-Q_1	.683	.617	.650	.787	.712	.749
		$S-\chi^2$	.378	.421	.399	.548	.575	.561
		χ^2	.762	.695	.728	.847	.776	.811
		G^2	.738	.666	.702	.825	.746	.785
60	500	*z(r)*	.948	.965	.957	.976	.984	.980
		z(*l*)	.934	.963	.948	.968	.982	.975
		Stone-Q_1	.767	.689	.728	.845	.761	.803
		$S-\chi^2$	.190	.269	.230	.356	.451	.404
		χ^2	.790	.703	.747	.853	.772	.813
		G^2	.753	.672	.713	.823	.745	.784
	1000	*z(r)*	.985	.997	.991	.992	.999	.996
		z(*l*)	.977	.996	.986	.987	.999	.993
		Stone-Q_1	.874	.740	.807	.917	.796	.857
		$S-\chi^2$	.484	.544	.514	.650	.704	.677
		χ^2	.890	.763	.827	.922	.811	.866
		G^2	.880	.752	.816	.915	.803	.859

表 6 低质量测验下生成模型为 DINA 的统计检验力

			α=.01		平均值	α=.05		平均值
L	*N*	统计量	DINO	ACDM		DINO	ACDM	
30	500	*z(r)*	.264	.022	.143	.444	.054	.249
		z(*l*)	.203	.012	.108	.341	.041	.191
		Stone-Q_1	.159	.075	.117	.325	.213	.269
		$S-\chi^2$	.128	.040	.084	.270	.122	.196
		χ^2	.213	.141	.177	.378	.311	.344
		G^2	.167	.121	.144	.313	.279	.296
	1000	*z(r)*	.499	.044	.272	.689	.118	.403
		z(*l*)	.390	.026	.208	.605	.090	.348
		Stone-Q_1	.363	.200	.281	.515	.378	.447
		$S-\chi^2$	.284	.069	.177	.446	.162	.304
		χ^2	.448	.396	.422	.589	.566	.577
		G^2	.415	.373	.394	.552	.540	.546
60	500	*z(r)*	.517	.029	.273	.724	.081	.403
		z(*l*)	.384	.014	.199	.606	.053	.330
		Stone-Q_1	.513	.239	.376	.609	.448	.528
		$S-\chi^2$	.106	.016	.061	.235	.057	.146
		χ^2	.551	.300	.426	.621	.486	.554
		G^2	.518	.270	.394	.600	.440	.520
	1000	*z(r)*	.865	.081	.473	.953	.201	.577
		z(*l*)	.778	.044	.411	.922	.135	.529
		Stone-Q_1	.631	.520	.575	.659	.686	.672
		$S-\chi^2$	.285	.039	.162	.454	.126	.290
		χ^2	.636	.592	.614	.669	.716	.693
		G^2	.634	.562	.598	.663	.688	.676

指标对于 2 种拟合（假）模型的平均统计检验力。

从表 4 可知，当生成模型为 ACDM 时，在低质量的测验条件下，$z(r)$ 和 $z(l)$ 的平均统计检验力明显高于其他方法，但此时，$z(r)$ 和 $z(l)$ 的最高统计检验力仍未达到 .8。当 L= 60 时，每个题目拟合方法的统计检验力也随之提高，在 N=1000 时，$z(r)$ 和 $z(l)$ 的平均统计检验力已达到了 .8，而在 α = .05 时，接近于 .9。

表 5 的结果显示，在高质量的条件中，$z(r)$ 和 $z(l)$ 的统计检验力最高，S-χ^2 的统计检验力最低。与表 4 低质量测验的条件相比，在同等条件下，高质量测验的统计检验力有很大的提高，例如，在 N=500 和 L=30 时，$z(r)$ 和 $z(l)$ 的平均统计检验力已经超过了 .85，在 α = .05 时，已经超过了 .9。

表 6 显示，生成模型为 DINA 时，在低质量测验中，当用 DINO 拟合数据时，$z(r)$ 和 $z(l)$ 具有最高的统计检验力，其中 $z(r)$ 的统计检验力明显高于 $z(l)$。但当拟合模型是 ACDM 时，χ^2 和 G^2 具有最高的统计检验力。因此，就平均统计检验力而言，$z(r)$ 和 $z(l)$ 由于在拟合模型 ACDM 中表现不佳，导致整体的平均统计检验力要低于 χ^2 和 G^2。

表 7 高质量测验下生成模型为 DINA 的统计检验力

L	N	统计量	α=.01 DINO	α=.01 ACDM	平均值	α=.05 DINO	α=.05 ACDM	平均值
30	500	$z(r)$	.956	.417	.687	.976	.545	.761
		$z(l)$	.911	.230	.571	.966	.356	.661
		Stone-Q_1	.597	.539	.568	.639	.684	.661
		S - χ^2	.381	.091	.236	.545	.186	.365
		χ^2	.607	.633	.620	.651	.737	.694
		G^2	.593	.570	.581	.635	.692	.664
	1000	$z(r)$	.994	.673	.833	.998	.780	.889
		$z(l)$	.990	.519	.755	.997	.673	.835
		Stone-Q_1	.648	.718	.683	.694	.813	.754
		S - χ^2	.674	.264	.469	.785	.456	.620
		χ^2	.653	.800	.726	.702	.862	.782
		G^2	.655	.774	.714	.700	.852	.776
60	500	$z(r)$	.999	.633	.816	1.000	.746	.873
		$z(l)$	.986	.410	.698	.995	.549	.772
		Stone-Q_1	.657	.742	.699	.695	.813	.754
		S - χ^2	.450	.057	.254	.610	.117	.363
		χ^2	.662	.768	.715	.696	.825	.761
		G^2	.647	.684	.666	.678	.775	.727
	1000	$z(r)$	1.000	.832	.916	1.000	.901	.950
		$z(l)$	1.000	.689	.845	1.000	.794	.897
		Stone-Q_1	.687	.854	.770	.730	.895	.813
		S - χ^2	.709	.195	.452	.789	.321	.555
		χ^2	.693	.872	.782	.734	.905	.819
		G^2	.691	.847	.769	.733	.891	.812

表 7 的结果显示，在高质量的测验中，当拟合模型是 DINO 时，$z(r)$ 和 $z(l)$ 的效果相似，但当拟合模型是 ACDM 时，$z(r)$ 的统计检验力要远好于 $z(l)$。在 L=30 的条件下，$z(r)$ 仍然不能有效地区分 DINA 和 ACDM，但当 L=60 时，$z(r)$ 的统计检验力有大幅度的提高。

表 8 和 9 的结果表明，当 DINO 为生成模型时，$z(r)$ 和 $z(l)$ 在区分 DINO 和 ACDM 时的效果并不理想。在低质量测验中，当拟合模型是 DINA 时，$z(r)$ 具有最高的统计力，而当用 ACDM 拟合数据时，$z(r)$ 和 $z(l)$ 的表现都很差。在高质量测验中，当拟合模型是 DINA 时，$z(r)$ 和 $z(l)$ 的统计检验力都超过了 0.9，但当用 ACDM 拟合时，$z(r)$ 的统计检验力要高于 $z(l)$。因此，就平均统计检验力而言， 在低质量条件下，Stone-Q_1，χ^2 和 G^2 的表现更好，而在高质量条件下，$z(r)$ 的平均统计检验力达到最高。

4 实测数据分析

为了验证题目拟合方法在实证数据的效果，现通过一个实证数据的分析来加以说明。实证数据取自 Chen 和 de la Torre（2014）分析过的一个 PISA

表 8 低质量测验下生成模型为 DINO 的统计检验力

			α=.01		平均值	α=.05		平均值
L	N	统计量	DINA	ACDM		DINA	ACDM	
30	500	$z(r)$	.260	.010	.135	.440	.046	.243
		z(l)	.190	.008	.099	.336	.036	.186
		Stone-Q_1	.172	.074	.123	.333	.214	.273
		S-χ^2	.128	.040	.084	.259	.118	.188
		χ^2	.209	.149	.179	.376	.324	.350
		G^2	.155	.130	.143	.289	.288	.288
	1000	$z(r)$	.497	.041	.269	.694	.110	.402
		z(l)	.392	.027	.209	.603	.077	.340
		Stone-Q_1	.407	.196	.301	.550	.375	.462
		S-χ^2	.271	.068	.170	.438	.181	.309
		χ^2	.471	.399	.435	.618	.577	.597
		G^2	.435	.386	.410	.588	.555	.572
60	500	$z(r)$	.512	.024	.268	.719	.080	.399
		z(l)	.386	.012	.199	.608	.049	.328
		Stone-Q_1	.506	.237	.372	.600	.438	.519
		S-χ^2	.104	.014	.059	.232	.054	.143
		χ^2	.541	.311	.426	.609	.489	.549
		G^2	.501	.281	.391	.583	.441	.512
	1000	$z(r)$	.848	.085	.467	.950	.191	.570
		z(l)	.767	.052	.409	.909	.134	.522
		Stone-Q_1	.624	.523	.574	.654	.688	.671
		S-χ^2	.291	.044	.167	.461	.134	.298
		χ^2	.636	.601	.619	.666	.722	.694
		G^2	.633	.568	.601	.665	.699	.682

表 9 高质量测验下生成模型为 DINO 的统计检验力

			α=.01		平均值	α=.05		平均值
L	N	统计量	DINA	ACDM		DINA	ACDM	
30	500	$z(r)$	.955	.389	.672	.978	.534	.756
		z(l)	.907	.202	.554	.958	.343	.650
		Stone-Q_1	.604	.513	.559	.645	.661	.653
		S-χ^2	.382	.090	.236	.556	.195	.375
		χ^2	.614	.615	.615	.653	.724	.688
		G^2	.596	.544	.570	.642	.669	.655
	1000	$z(r)$	.994	.687	.840	.998	.800	.899
		z(l)	.990	.510	.750	.997	.682	.839
		Stone-Q_1	.692	.737	.715	.704	.830	.767
		S-χ^2	.647	.272	.460	.784	.451	.617
		χ^2	.663	.812	.737	.711	.880	.796
		G^2	.656	.786	.721	.710	.860	.785
60	500	$z(r)$	1.000	.627	.814	1.000	.736	.868
		z(l)	.989	.389	.689	.999	.545	.772
		Stone-Q_1	.661	.716	.689	.705	.801	.753
		S-χ^2	.455	.052	.254	.614	.109	.361
		χ^2	.670	.757	.713	.705	.817	.761
		G^2	.648	.687	.667	.688	.770	.729
	1000	$z(r)$	1.000	.835	.918	1.000	.910	.955
		z(l)	1.000	.705	.853	1.000	.809	.905
		Stone-Q_1	.690	.864	.777	.742	.900	.821
		S-χ^2	.711	.196	.453	.796	.329	.562
		χ^2	.699	.877	.788	.746	.906	.826
		G^2	.698	.855	.777	.740	.894	.817

表 10 测验水平的拟合检验

模型	Deviance	AIC	BIC	CAIC	SABIC
DINA	30608	31222	32757	33064	31782
DINO	30728	31342	32877	33184	31902
ACDM	29952	30674	32479	32840	31332

阅读数据。该数据包括 1095 名学生在 26 个题目的作答，测验考察了 8 个属性，关于该数据的更详细介绍可参考 Chen 和 de la Torre（2014）。

首先，通过测验水平的拟合检验从 DINA，DINO 和 ACDM 中选择一个相对拟合最优的 CDM，表 10 显示了测验水平的拟合检验结果。正如表 10 所示，ACDM 模型的 4 个整体拟合指标值都是最小，表明 ACDM 是拟合更好的模型。因此，后续的分析采用了 ACDM 模型。

然后，进行了题目水平的拟合检验，模拟实验的结果显示当生成模型是 ACDM 时，在样本容量为 1000 时，综合考虑一类错误率和统计检验力，$z(r)$ 和 $z(l)$ 的效果相似，并且是所有方法中表现最好的。通过比较 2 个指标的实际分析效果发现，它们筛选出的拟合不佳题目是一致的。因此，这 2 种题目拟合指标可任选一种报告，表 11 汇总了 $z(l)$ 指标的题目拟合检验结果，其中 p 值是经过 Holm 法调整后的结果。表 11 的结果显示在 $\alpha = .01$ 水平下，大部分题目拟合良好，只有包括 13、16 和 17 题在内的 3 个题目显示拟合不佳。

表 11 题目水平的拟合检验（$z[l]$ 方法）

题目	p	题目	p
1	.063	14	.415
2	.149	15	.045
3	.024	16	.005
4	1.000	17	.004
5	1.000	18	.431
6	.131	19	.015
7	.551	20	.441
8	1.000	21	.479
9	.376	22	.045
10	1.000	23	.717
11	.163	24	1.000
12	.012	25	.034
13	.001	26	.376

5 讨论

当前，已有研究将 IRT 中常用的题目拟合统计量拓展到 CDM 中，例如基于卡方类的统计量和基于题目对的统计量。然而，仍没有研究比较过这些题目拟合指标在 CDM 中的综合效果。本研究系统比较了这些题目拟合指标在不同实验条件下的表现。实验结果显示，就所犯一类错误率而言，在所有条件下，$z(r)$ 和 $z(l)$ 的表现稳定，且效果更好。就统计检验力而言，当生成模型为 ACDM 时，在所有条件下 $z(r)$ 和 $z(l)$ 都有最高的平均统计检验力。当生成模型为 DINA 或 DINO 时，在低质量测验中，χ^2 和 G^2 的统计检验力更高；而在高质量测验中，$z(r)$ 有最高的统计检验力。因此，综合一类错误率和统计检验力来考虑，在实际数据分析中，如果数据整体拟合 ACDM，LLM 或 R-RUM，这类模型具有相似的计量性能，建议选择 $z(r)$ 和 $z(l)$ 来评价题目拟合；当数据整体拟合 DINA 或 DINO 模型时，在低质量测验中，建议选择 χ^2 和 G^2；若在高质量测验中，建议选择 $z(r)$ 统计量。

本研究的模拟实验固定属性个数为 5 个，但在实际应用中，属性个数可能会更多，例如，Xi 等人（2020）将认知诊断模型应用于分裂型人格的诊断中，并将 DSM-V 对分裂型人格障碍的 9 个诊断标准视作认知属性。未来的研究应关注属性数量对题目拟合方法的影响。此外，在平常的测验中，经常会出现一些异常作答的被试。对于这类被试而言，测验结果并不能恰当地反应他们真实的水平。因此，进行被试拟合（person fit）评估也是认知诊断测验中的一个重要环节，它可以帮助验证个别学生的诊断结果（Cui & Li, 2015）。当前，对于认知诊断模型中被试拟合检验的研究还很少，未来有待开展更多的研究。

参考文献

涂冬波 , 张心 , 蔡艳 , 戴海琦 . (2014). 认知诊断模型 – 资料拟合检验统计量及其性能 . *心理科学 , 37*(1), 205–211.

Acevedo–Mesa, A., Tendeiro, J., Roest, A., Rosmalen, J., & Monden, R. (2020). Improving the measurement of functional somatic symptoms with item response theory. *Journal of Psychosomatic Research, 133,* Article 110009.

Chalmers, R. P., & Ng, V. (2017). Plausible–value imputation statistics for detecting item misfit. *Applied Psychological Measurement, 41*(5), 372–387.

Chen, J. S., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large–scale assessment data: The case of the programme for international student assessment in reading. *Psychology, 5*(18), 1967–1978.

Chen, J. S., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Cui, Y., & Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement, 39*(3), 223–238.

de la Torre, J., & Douglas, J. A. (2004). Higher–order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, 51*(4), 281–296.

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2019). Development of a computerized adaptive test for anxiety based on the Dutch–Flemish version of the PROMIS item bank. *Assessment, 26*(7), 1362–1374.

Gao, X. L., Wang, D. X., Cai, Y., & Tu, D. B. (2020). Cognitive diagnostic computerized adaptive testing for polytomously scored items. *Journal of Classification, 37*(3), 709–729.

Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias–corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics, 45*(3), 251–273.

Li, H. L., Kim, M. K., & Xiong, Y. (2020). Individual learning vs. interactive learning: A cognitive diagnostic analysis of MOOC students' learning behaviors. *American Journal of Distance Education, 34*(2), 121–136.

Liu, Y., & Maydeu–Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research, 49*(4), 354–371.

Ma, W. C., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200–217.

Orlando, M., & Thissen, D. (2000). Likelihood–based item–fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*(3), 506–532.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item–fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*(8), 614–631.

Su, S. Y., Wang, C., & Weiss, D. J. (2021). Performance of the S–χ^2 statistic for the multidimensional graded response model. *Educational and Psychological Measurement, 81*(3), 491–522.

Wang, C., Shu, Z., Shang, Z. R., & Xu, G. J. (2015). Assessing item–level fit for the DINA model. *Applied Psychological Measurement, 39(7*), 525–538.

Xi, C. Q., Cai, Y., Peng, S. W., Lian, J., & Tu, D. B. (2020). A diagnostic classification version of Schizotypal Personality Questionnaire using diagnostic classification models. *International Journal of Methods in Psychiatric Research, 29*(1), Article e1807.

Zhang, X., Wang, C., & Tao, J. (2018). Assessing item–level fit for higher order item response theory models. *Applied Psychological Measurement, 42*(8), 644–659.

Evaluation of Item-Level Fit in Cognitive Diagnosis Model

Gao Xuliang, Wang Fang, Xia Linpo, Hou Minmin

（School of Psychology Guizhou Normal University, Guiyang, 550025）

(Mental Health Education and Counseling Center, Guizhou Normal University, Guiyang, 550025）

Abstract The goal of cognitive diagnosis model (CDM) is to classify participants into potential categories with different attribute patterns, which provide diagnostic information about whether the student has mastered a set of skills or attributes. Compared with single-dimensional item response models (e.g., item response models), CDM provides a more detailed assessment of the strengths and weaknesses of students. Although CDM was originally developed in the field of educational evaluation, it has now been used to evaluate other types of structures, such as psychological disorders and context-based ability assessment. As with any model-based evaluation, a key step in implementing the CDM is to check the model data fit, that is, the consistency between model predictions and observed data. Only when the model fits the data, the estimated model parameters can be reliably explained. Item fit is used to evaluate the fit of each item with the model, which helps to identify abnormal items. Deleting or modifying these items will improve the overall model data fit for the entire test.

At present, some commonly used item fit statistics in IRT have been extended to CDM. However, there is no research system to compare the comprehensive performance of these item fit indicators in CDM. In this study, we compared the performance of χ^2, G^2, S-χ^2, $z(r)$, $z(l)$, and Stone-Q_1 in the CDM. This study investigated the Type I error rate and power of the above item fit statistics through a simulation study. The factors manipulated include sample size (N=500, 1000), generating model (DINA, DINO, and ACDM), fitting model (DINA, DINO, and ACDM), test length (30 and 60), test quality (high and low), and significance level (.01 and .05). The test examined five attributes. For high-quality and low-quality tests, the guess parameters and slipping parameters of the three generating models are randomly extracted from uniform distributions U(.05, .15) and U(.15, .25), respectively.

The simulation results showed that, in terms of the Type I error, $z(r)$ and $z(l)$ performed best under all conditions. In terms of statistical test power, when the generating model was ACDM, $z(r)$ and $z(l)$ had the highest average power under all conditions. When the generating model was DINA or DINO, in the low-quality test, the power of $\chi^2$and G^2 was higher; and in the high-quality test, $z(r)$ had the highest power. In short, combining the performance of the Type I error and power, if the data fit A-CDM, $z(r)$, and $z(l)$performed best; when the data fit the DINA or DINO model, in low-quality test, χ^2, and G^2 performed the best; however, in high-quality tests, the $z(r)$ performed better among all methods.

This study only investigated the condition that the number of attributes is 5, and the actual test may measure more attributes. Therefore, future research should focus on the influence of the number of attributes. Lastly, the person fit assessment is also an important step in the cognitive diagnostic test, which can help identify the abnormal responses of individual students. More studies on the person fit in cognitive diagnosis model are needed.

Key words CDM, item fit, Type I error rate, power