认知诊断模型的比较及其应用研究：饱和模型、简化模型还是混合方法*

高旭亮 汪大勋 蔡 艳 涂冬波**
（江西师范大学心理健康教育研究中心，江西师范大学心理学院，南昌，330022）

摘 要 GDINA 是一个饱和认知诊断模型，Wald 检验被用于在题目水平上检验 GDINA 是否可以被简化模型（如 DINA, DINO, ACDM 和 CRUM）替代，并为测验的每一个题目选择一个最恰当的 CDM（简称混合方法）。选择合适的 CDMs 是进行诊断评估的一个关键步骤，通过 Monte Carlo 模拟实验，比较了不同的测验情境下，GDINA、简化 CDMs 和混合方法在测验整体拟合指标、模式判准率和项目参数估计的返真性等效果，研究发现混合方法的整体表现是最好的，其次是 GDINA，最后是简化 CDMs。

关键词 GDINA 饱和模型 Wald 检验 简化模型

1 引言

近年来，认知诊断评估测验逐渐在教育评价和心理疾病的临床诊断中得到了初步的应用（de la Torre, van der Ark, & Rossi, 2015; Jang, 2009; Kim, 2015; Zhang, 2013），对 CDMs 的研究逐渐受到学者们的重视，并且基于不同的理论假设提出了一系列 CDMs。这些 CDMs 中既有约束条件非常少、适用面相对较广的饱和模型，也有约束条件较多的简化模型。比较常见的简化模型有 DINA（Haertel, 1989）、DINO（Templin & Henson, 2006）、ACDM（de la Torre, 2011）、CRUM（Hartz, 2002）等；饱和模型有 GDINA（de la Torre, 2011），GDM（von Davier, 2008）和 LCDM（Henson, Templin, & Willse, 2009），饱和模型在一定的约束条件下可以简化为简化模型，因此简化模型是饱和模型的特例。

简化模型的优点是项目参数少，参数的估计要达到稳定状态所需样本相对要少，诊断结果更加直观和易于解释，但其缺点是模型假设比较严格，这在一定程度上限制了简化模型的应用。与简化模型相比，饱和模型的理论假设比较宽松，对属性之间的作用并未严格限制，因此，饱和模型的适用面更广。但饱和模型的缺点是模型较复杂且不易于解释，待估计的项目参数较多，需要更大的样本才有可能准确估计项目参数（高旭亮 , 涂冬波 , 2017）。另外，有大量研究（de la Torre & Lee, 2013; Li, Hunter, & Lei, 2016; Ma, Iaconangelo, & de la Torre, 2016)建议，当简化模型和饱和模型在模型拟合度相当的条件下，优先选择简化模型。Rojas，de la Torre 和 Olea（2012）研究发现与 GDINA 相比，使用恰当的简化 CDMs 会有更高的模式判准率。

饱和模型和简化模型各有优缺点，那么在实际应用中，是否有可能在一次认知诊断分析中同时使用饱和模型和简化模型，即有的项目使用饱和模型而有的项目使用简化模型，本文将这种做法称为混合方法（mixed method），从而充分发挥两类模型的优点。de la Torre 和 Lee（2013）以及 Ma, Iaconangelo 和 de la Torre（2016）通过实证数据的分析发现，一份测验中不同题目会拟合不同的 CDMs。de la Torre, van der Ark 和 Rossi（2015）提出在 GDINA 框架之下，在题目水平之上通过 Wald 统计量检验饱和模型是否能被简化模型所替代，从而为每一个题目选用最合适的 CDM；此时，有的题目可能选用饱和模型（如 GDINA）而有的题目可能选用简化模型（如 DINA、DINO、CRUM 或 ACDM 等），这也正是混合方法的思路。显然，理论上混合方法同时具有饱和模型和简化模型的双重优点，在实际的应用中具有较好的应用前景，值得进一步研究。

* 本研究国家自然科学基金 (31660278, 31760288)、江西省高校人文社科项目 (XL1507，XL1508)、江西省社科规划项目 (17JY12) 和江西省高等院校教学改革研究课题（JXJG–15–2–26）的资助。

** 通讯作者：涂冬波。E–mail: tudongbo@aliyun.com

DOI:10.16719/j.cnki.1671–6981.20180333

如何在诊断评估测验中从这三类模型中选择合适的 CDMs 值得研究。查阅国内外相关文献，关于三类模型的比较与选用研究非常薄弱。涂冬波，蔡艳和戴海琦（2013）讨论了关于简化模型的比较与选用，但并不涉及饱和模型和混合方法；Ma, Iaconangelo 和 de la Torre（2016）初步探讨了简化模型为真模型时，混合方法和饱和模型去拟合数据时的效果，但没有考虑不同真模型下模型选用的效果及其比较，也没有讨论不同测验 Q 矩阵对模型选用的影响等。这为本文的研究提供了基础。鉴于 CDMs 的选用在实际运用中的重要性以及当前国内外对一领域的研究相对薄弱，本研究拟在前人的基础上，进一步探讨在不同测验情景下三类 CDMs 的选用及其效果，并对不同情境中的如何选择合适的 CDMs 提出具体的使用建议，为实际应用者在 CDMs 的选用上提供借鉴。

2 混合方法

混合方法的思路是 de la Torre, van der Ark 和 Rossi（2015）提出，它的目的是为每一个题目选择适当的 CDM，分为两步：第一步，所有题目用 GDINA 来估计参数；第二步，应用 Wald 检验在题目水平上比较 DINA、DINO、ACDM 和 CRUM 是否可以取代 GDINA，显著水平设为 0.05；如果 Wald 检验选出了 2 个以上的简化模型，则选择其中参数较少的模型。根据以上步骤为每一题选择一个最优的 CDM，并重新估计题目参数。

3 研究方案

3.1 实验设计

实验有五个因素：测验人数，题目质量，产生模型（真模型），匹配模型（fitted model），Q 矩阵的复杂度。

（1）测验人数和题目质量：500 和 1000 人；题目质量分为高、低水平，模拟方法参考 Ma 和 de la Torre（2016）的做法，高质量的题目固定 $P(1)$=.1，$P(0)$=.1，$P(1)$ 表示掌握了题目考察的所有属性的答对概率是 .9，$P(0)$ 代表被试没有掌握题目考察的任意属性而猜对的概率，而其他可能的 KS 答对概率从均匀分布 U(.1, .9) 中随机生成；低质量时 $P(1)$=.3，$P(0)$=.3 。

（2）真模型和匹配模型：GDINA，Mixed，DINA，DINO，ACDM，CRUM。

（3）Q 矩阵：简单 Q 矩阵有 10 题分别考察了一个属性，另外各有 10 题分别考察了 2 个属性和 3 个属性，复杂 Q 矩阵则将简单 Q 矩阵最后考察 3 个属性的 5 个题目替换成考察 4 个属性。

3.2 Monte Carlo 实验

（1）被试和项目参数的模拟：被试 KS 假设服从均匀分布，项目参数模拟参见上述题目质量的模拟方法。

（2）作答数据的模拟：真模型是 GDINA 时：首先用 GDINA 模拟 个被试作答 题数据，然后将作答数据交给 6 个匹配模型重新估计参数。

真模型是 Mixed 时：将 个题目随机分为 6 份，随机分配给 GDINA, DINA, DINO, ACDM 和 CRUM 模拟 个人的作答数据，并用 6 种匹配模型重新估计。

简化模型为真模型：为了便于说明此处仅以 CRUM 为例加以说明，用 CRUM 模拟 个被试作答题数据，然后将作答数据交给 6 种匹配模型重新估计。

（3）重复以上模拟过程 100 次，本研究的参数估计使用了 R 软件（3.4.0 版本）中的 GDINA 程序包（Ma & de la Torre, 2016）。

4 评价指标

4.1 测验拟合度指标

$$AIC=-2\ln(L)+2\times K \tag{1}$$

$$MAIC=\frac{1}{R}\sum_{r=1}^{R}AIC_r \tag{2}$$

其中 L 是似然值，K 是待估计的参数个数，AIC 指标越小代表模型拟合的越好，$MAIC$ 是计算 R=100 次实验的整体拟合指标 AIC 的均值。

4.2 属性诊断正确率指标

$$PCV=\frac{\sum_{r=1}^{R}\sum_{i=1}^{N}I^{(r)}\left[a_i=\hat{a}_i\right]}{N\times R} \tag{3}$$

N 表示被试数，R=100，$I^{(r)}\left[a_i=\hat{a}_i\right]$表示估计的 KS 和真实 KS 在第 r 次实验中是否相同，PCV 指标越大意味着被试参数估计的越准。

4.3 项目参数的估计精度指标

$$RMSE=\sqrt{\frac{\sum_{r=1}^{R}\sum_{c=1}^{2^K}\sum_{j=1}^{J}\left[\hat{P}^{(r)}\left(X_j=b\mid a_c\right)-P^{(r)}\left(X_j=b\mid a_c\right)\right]^2}{J\times 2^K\times R}} \tag{4}$$

其计算借鉴 Ma 和 de la Torre（2016）的做法，

J 是题目数，K 是属性个数，R=100，$\hat{P}^{(r)}\left(X_j=b\mid a_c\right)$ 和 $P^{(r)}\left(X_j=b\mid a_c\right)$ 分别代表知识状态 a_c 答对第 j 题的估计和真实作答概率。例如，考察 5 个属性时，可能的 KS 有 32 种，每一题有 32 种可能的作答概率，并将每一题估计的 32 种作答概率值视作项目参数估计值。

表 1　模型资料整体拟合指标（MAIC）

Q 矩阵	样本量	真模型	匹配模型											
			低质量						高质量					
			GDINA	Mixed	DINA	DINO	ACDM	CRUM	GDINA	Mixed	DINA	DINO	ACDM	CRUM
简单	500	GDINA	**20434**	20420	20510	20540	20454	20446	**15970**	15963	17688	17745	16401	16345
		Mixed	20368	**20350**	20439	20435	20400	20399	15125	**15080**	16864	17008	15977	15750
		DINA	19896	19857	**19849**	20020	19954	19930	13103	13026	**13026**	16180	15099	13604
		DINO	19954	19960	20098	**19947**	20032	20037	13235	13148	16400	**13148**	15262	15533
		ACDM	20479	20458	20611	20619	**20483**	20480	15812	15767	17913	17895	**15776**	15842
		CRUM	20460	20459	20591	20611	20465	**20459**	15723	15663	17358	17994	15961	**15664**
		Mean	20265	20251	20350	20362	20298	20292	14828	14775	16541	16662	15746	15456
	1000	GDINA	**40790**	40755	40932	40945	40796	40796	**31842**	31836	35302	35434	32651	32632
		Mixed	40571	**40539**	40811	40822	40592	40585	29755	**29704**	34393	33322	31740	31708
		DINA	39519	39446	**39444**	39784	39668	39606	25785	25718	**25720**	32223	29852	26738
		DINO	39734	39663	39991	**39667**	39868	39897	26316	26221	32856	**26221**	30350	30965
		ACDM	40783	40738	41090	41094	**40751**	40753	31643	31587	36096	36017	**31589**	31835
		CRUM	40764	40723	41018	41037	40725	**40725**	30909	30845	34499	35639	31334	**30845**
		Mean	40360	40311	40548	40558	40400	40393	29375	29319	33144	33143	31253	30787
复杂	500	GDINA	**20397**	20394	20521	20500	20438	20447	**16451**	16448	18310	18083	16928	16960
		Mixed	20332	**20331**	20432	20450	20361	20352	15730	**15728**	17752	17739	16602	16400
		DINA	19669	19676	**19632**	19760	19697	19696	12891	12797	**12746**	15380	14633	13404
		DINO	19653	19666	19793	**19627**	19734	19746	12915	12904	15470	**12801**	14680	14844
		ACDM	20571	20579	20731	20723	**20573**	20574	17044	16958	19101	19038	**16938**	17086
		CRUM	20521	20517	20673	20688	20507	**20506**	16583	16473	18240	18652	16744	**16473**
		Mean	20190	20199	20297	20291	20218	20220	15269	15227	16937	16949	16088	15861
	1000	GDINA	**40736**	40733	40904	40848	40764	40769	**32792**	32784	36484	36309	33884	33976
		Mixed	40492	**40479**	40679	40674	40555	40557	30771	**30654**	34587	34288	32631	32287
		DINA	39137	39121	**39050**	39301	39247	39198	25376	25246	**25246**	30787	29263	26441
		DINO	39280	39281	39524	**39230**	39408	39431	26000	25865	31329	**25873**	29803	30071
		ACDM	41012	40986	41289	41275	**40995**	40999	33338	33229	37856	37751	**33228**	33515
		CRUM	40932	40912	41254	41258	40916	**40912**	32596	32485	36742	37326	33093	**32492**
		Mean	40265	40252	40450	40431	40314	40311	30146	30044	33707	33722	31984	31464

5　实验结果

表 1 的行表示真模型，列对应 6 个匹配模型重新估计值，6 个真模型对应 6 个匹配模型，对角线元素代表真模型的拟合指标，Mean 行是该矩阵按列求平均值的结果，代表了每个匹配模型在 6 种真模型条件下的平均拟合指标。表 2、3 的结构和表 1 相同。

真模型是 GDINA 和 Mixed 时，Mixed 的拟合度最好，其次是 GDINA；而当真模型是简化模型时，Mixed 和真模型拟合度表现相似。当真模型是 DINA 时，DINO 的拟合度最差，而当真模型时 DINO 时，DINA 的是拟合最差的模型，当真模型是 ACDM 或 CRUM 时，尤其在低质量题目时，这两个模型之间的拟合度几乎没有差异，导致这种结果的原因是由模型之间的计量特点所决定的，DINA 和 DINO 的模型假设正好相反，而 ACDM 和 CRUM 模型具有相似的模型假设。

简单 Q 阵，真模型是 Mixed 时，Mixed 的判准率最高，其次是 GDINA，CRUM，ACDM，DINA 和 DINO；真模型是 GDINA，CRUM 和 ACDM 时，Mixed 和这三种真模型的判准率很相似。当真模型是 DINA 或者 DINO 时，真模型的表现最好，其次是 Mixed。DINA 是真模型时 DINO 的表现最差，反之亦然。复杂 Q 阵，低质量题目时，整体的模式判准率很低，当真模型是 GDINA 时，ACDM 和 CRUM 的效果最好，其次是 Mixed、GDINA、DINA 和 DINO；当真模型是 Mixed 时，Mixed 和 ACDM，CRUM 的效果相似；真模型是 4 种简化模型时，真模型的表现最好，随着题目质量的进一步

表 2 属性模式判准率指标（PCV）

Q 矩阵	样本量	真模型	匹配模型											
			低质量						高质量					
			GDINA	Mixed	DINA	DINO	ACDM	CRUM	GDINA	Mixed	DINA	DINO	ACDM	CRUM
简单	500	GDINA	**.207**	.224	.201	.154	.240	.242	**.896**	.895	.659	.495	.868	.876
		Mixed	.225	**.231**	.171	.155	.213	.221	.917	**.919**	.680	.608	.851	.883
		DINA	.217	.257	**.268**	.085	.158	.197	.899	.912	**.912**	.318	.662	.858
		DINO	.240	.240	.076	**.326**	.122	.118	.920	.920	.277	**.920**	.672	.572
		ACDM	.196	.233	.131	.123	**.293**	.269	.885	.885	.334	.265	**.895**	.880
		CRUM	.206	.246	.139	.117	.255	**.251**	.883	.898	.525	.304	.869	**.899**
		Mean	.214	.239	.164	.160	.214	.216	.901	.905	.565	.485	.803	.828
	1000	GDINA	**.271**	.280	.173	.156	.281	.281	**.905**	.905	.680	.659	.872	.875
		Mixed	.264	**.276**	.151	.146	.276	.277	.924	**.926**	.765	.586	.836	.870
		DINA	.345	.357	**.357**	.095	.226	.270	.907	.908	**.908**	.543	.745	.883
		DINO	.303	.359	.081	**.359**	.245	.224	.930	.930	.327	**.931**	.738	.712
		ACDM	.287	.303	.141	.151	**.310**	.309	.897	.901	.456	.452	**.903**	.901
		CRUM	.260	.281	.164	.134	.294	**.295**	.900	.904	.574	.293	.884	**.905**
		Mean	.288	.309	.176	.174	.272	.276	.911	.912	.618	.577	.829	.857
复杂	500	GDINA	**.185**	.198	.066	.143	.211	.194	**.808**	.810	.294	.268	.744	.770
		Mixed	.171	**.171**	.103	.087	.176	.183	.868	**.868**	.331	.229	.744	.755
		DINA	.122	.153	**.235**	.044	.081	.131	.798	.798	**.802**	.054	.220	.742
		DINO	.233	.255	.028	**.311**	.092	.091	.808	.808	.066	**.809**	.187	.156
		ACDM	.151	.205	.110	.148	**.220**	.220	.834	.834	.230	.219	**.837**	.837
		CRUM	.187	.201	.125	.097	.228	**.225**	.788	.814	.291	.184	.758	**.817**
		Mean	.175	.197	.111	.138	.168	.174	.817	.822	.336	.294	.582	.679
	1000	GDINA	**.185**	.196	.083	.111	.223	.213	**.854**	.854	.205	.223	.786	.776
		Mixed	.173	**.209**	.084	.129	.219	.222	.878	**.881**	.272	.296	.766	.792
		DINA	.173	.221	**.293**	.038	.117	.158	.830	.836	**.836**	.054	.322	.781
		DINO	.186	.195	.048	**.428**	.211	.164	.763	.855	.063	**.879**	.461	.483
		ACDM	.181	.223	.083	.086	**.285**	.285	.857	.862	.233	.231	**.863**	.853
		CRUM	.195	.233	.109	.089	.256	**.266**	.875	.879	.345	.163	.830	**.879**
		Mean	.182	.213	.117	.147	.219	.218	.843	.861	.326	.308	.671	.760

提高，模式判准率整体得到了提高。

表 3 可以发现，题目是低质量时，不管真模型是 GDINA 或者 Mixed，Mixed 的估计误差是最小的，而当真模型是 4 种简单模型时，真模型的估计误差最小，其次是 Mixed、GDINA。当题目是高质量时，真模型的 RMSE 指标是最小的，其次也是 Mixed。

总体来看，拟合最优的模型对应的模式判准率指标和估计误差也是最小的，当真模型是 GDINA 时，Mixed 的综合表现甚至略好于 GDINA，而当真模型是 Mixed 时，Mixed 的总体表现是最好的，其次是 GDINA，也就是说当一份测验中的题目可能拟合不同的模型时，选用 Mixed 方法是较理想的做法，其次可以选择 GDINA 模型。但当真模型是简化模型时，Mixed 的整体效果和真模型是很接近的，意味着当真模型是简化模型时（所有题目拟合同一个简化模型），优先选择简化模型（真模型）分析，其次也可以选用 Mixed 方法来分析，第三选择是 GDINA。

6 实证数据分析

为了进一步比较各模型在实际中的应用，我们采用了 Tatsuoka(1990)分数减法数据进行实证分析，分别从测验拟合、项目拟合和测量信度三个指标进行分析与比较，从而进一步探讨各模型在实际应用中的性能。

6.1 Tatsuoka 分数减法数据

Tatsuoka 分数减法数据最原始的版本可见 Tatsuoka (1990)。de la Torre（2011）对该数据 Q 矩阵重新标定了属性，被试数是 536，并将原来 20 题简化为 12 题，重新标定的 Q 矩阵共考查了 4 个属性。

6.2 各模型下测验整体拟合度分析及比较

常用的反应模型资料整体拟合指标有偏差（deviance）指标，AIC 和 BIC（Lei & Li，2016）。表 5 只列出了选用简单模型的题目，剩余题目选用了 GDINA 模型。

表 6 显示了分别用 GDINA、DINA、DINO、ACDM、CRUM 和混合方法拟合该数据的结果。

表 3 项目参数估计的均方根误差指标（RMSE）

Q 矩阵	样本量	真模型	匹配模型											
			低质量						高质量					
			GDINA	Mixed	DINA	DINO	ACDM	CRUM	GDINA	Mixed	DINA	DINO	ACDM	CRUM
简单	500	GDINA	**.132**	.102	.106	.127	.080	.084	**.040**	.042	.185	.210	.091	.088
		Mixed	.126	**.104**	.113	.122	.107	.099	.035	**.031**	.188	.189	.121	.106
		DINA	.139	.097	**.060**	.165	.135	.114	.032	.021	**.021**	.262	.183	.067
		DINO	.136	.107	.161	**.051**	.158	.167	.029	.019	.263	**.019**	.181	.209
		ACDM	.148	.123	.135	.152	**.063**	.070	.039	.034	.251	.258	**.027**	.048
		CRUM	.149	.106	.129	.165	.079	**.078**	.036	.027	.207	.264	.066	**.027**
		Mean	.138	.107	.117	.130	.103	.102	.035	.030	.186	.200	.112	.091
	1000	GDINA	**.131**	.104	.134	.135	.069	.073	**.036**	.038	.242	.240	.105	.110
		Mixed	.140	**.103**	.134	.130	.073	.074	.032	**.027**	.239	.208	.132	.119
		DINA	.141	.106	**.029**	.191	.146	.098	.029	.014	**.013**	.325	.224	.072
		DINO	.127	.105	.181	**.029**	.114	.120	.030	.014	.396	**.013**	.230	.303
		ACDM	.143	.092	.160	.159	**.052**	.058	.036	.026	.273	.276	**.024**	.049
		CRUM	.135	.091	.148	.168	.050	**.049**	.034	.025	.226	.286	.071	**.023**
		Mean	.136	.100	.131	.135	.084	.079	.033	.024	.231	.225	.131	.113
复杂	500	GDINA	**.169**	.129	.161	.120	.107	.114	**.052**	.054	.229	.230	.114	.118
		Mixed	.173	**.136**	.128	.131	.118	.118	.053	**.052**	.228	.230	.128	.118
		DINA	.182	.131	**.050**	.258	.196	.136	.042	.029	**.021**	.363	.260	.082
		DINO	.156	.124	.269	**.043**	.178	.181	.044	.039	.376	**.021**	.266	.309
		ACDM	.183	.110	.137	.129	**.075**	.074	.057	.044	.265	.271	**.038**	.059
		CRUM	.169	.111	.132	.141	.078	**.073**	.056	.040	.240	.293	.083	**.040**
		Mean	.172	.124	.146	.137	.125	.116	.050	.045	.227	.235	.148	.121
	1000	GDINA	**.131**	.104	.134	.135	.069	.073	**.036**	.039	.242	.240	.105	.110
		Mixed	.140	**.103**	.134	.130	.073	.074	.032	**.027**	.239	.208	.132	.119
		DINA	.141	.106	**.029**	.191	.146	.098	.029	.014	**.013**	.325	.224	.072
		DINO	.127	.105	.181	**.029**	.114	.120	.030	.014	.396	**.013**	.230	.303
		ACDM	.143	.092	.160	.159	**.052**	.058	.036	.026	.273	.276	**.024**	.049
		CRUM	.135	.091	.148	.168	.049	**.050**	.034	.025	.226	.286	.071	**.023**
		Mean	.136	.100	.131	.135	.084	.079	.033	.024	.232	.225	.131	.113

表 4 选用的模型

2	3	4	5	6	7	8	10	12
CRUM	DINO	CRUM	DINA	ACDM	DINO	DINA	CRUM	CRUM

表 5 模型资料拟合指标

	偏差	AIC	BIC
GDINA	5408.985	5650.985	6169.365
DINA	5589.869	5667.869	5834.950
DINO	5947.433	6025.433	6192.514
ACDM	5556.911	5674.911	5927.675
CRUM	5516.509	5634.509	5887.273
混合方法	5407.780	5525.780	5778.543

表 6 结果发现，单独以偏差和 AIC 指标来看，混合方法选择的模拟拟合最好，其次是 GDINA、CRUM、ACDM、DINA、DINO；从 BIC 指标来看，拟合度从高到低的顺序依次是：混合方法、DINA、CRUM、ACDM、GDINA 、DINO，这是由于与 AIC 相比 BIC 指标对于复杂模型的惩罚力度更大，更倾向于选择参数少的模型。总体来看混合方法在 3 个拟合指标的表现均是最优的。

6.3 各模型下测验项目拟合分析及比较

依据 Chen, de la Torre 和 Zhang（2013）提出的 log-odds ratio 统计量，计算了题目拟合指标。表 7 数据是该统计量对应的调整后的 p 值 (adjusted p-values)，该值小于 .05 意味着模型与题目拟合不良。

表 6 表明：混合方法选择的模型与题目拟合不良的题目数是最少的，其次是 GDINA，DINO 和 CRUM、ACDM、DINA。另外，第 1 题与所有的模

表 6 题目与模型的绝对拟合指标 (adjusted p-values)

题目	GDINA	DINA	DINO	ACDM	CRUM	混合方法
1	.001	.000	.000	.000	.000	.001
2	1.000	.000	.506	.000	.000	.586
3	.000	.001	.000	.000	.209	.000
4	.000	.227	.000	.015	.285	.586
5	.167	.000	.170	.000	.000	1.000
6	1.000	.000	1.000	.000	.000	.429
7	.000	.011	.000	1.000	.857	1.000
8	.000	.000	.000	.892	.318	.867
9	.000	.000	.000	.685	.339	1.000
10	1.000	.000	.440	.000	.000	.594
11	1.000	.000	.000	.000	.000	1.000
12	1.000	.000	.440	.000	.000	1.000
不拟合题目数	**6**	**11**	**7**	**9**	**7**	**2**

型都不拟合，第 3 题混合方法选择了 DINO，而从表 6 来看第 3 题与 CRUM 是拟合的，应该考虑选择 CRUM 模型来分析该题目。

6.4 各模型下测验整信度分析

Cui、Girel 和 Chang（2012）提出了认知诊断条件下的属性模式分类准确性（classification accuracy）分类一致性信度指标（classification consistency），表 7 是不同模型下的信度指标，可以看出基于混合方法分析下测验信度最高，可以最大程度上减少测验的测量误差。

表 7 认知诊断测验信度指标

模型	分类准确性	分类一致性
GDINA	.807	.804
DINA	.779	.769
DINO	.611	.629
ACDM	.667	.738
CRUM	.619	.811
混合方法	.879	.848

7 总结与讨论

目前为诊断测验选择模型的方法有很多，但并没有一种公认的方法是适合所有测验情境的。Lei 和 Li（2016）比较了多种选择模型的拟合指标，有从测验角度（相对拟合指标）包括：偏差、AIC、BIC 等指标，也有从题目的角度（绝对拟合指标）包括 RMSEA、MADcor、MADres、MADQ3、MX2 等指标，结果发现这些方法的效果取决于多种因素，如被试人数、题目质量、Q 矩阵是否正确等因素，即使在相同的测验情况下，不同指标选出的拟合模型可能并不相同。

混合方法通过 Wald 检验在题目水平上为每一个题目选择最优的模型。混合方法与传统做法不同，混合方法有它自身的优点，应用比较灵活，并不需要一个测验确定一个模型来分析，它为模型的选择以及诊断测验的开发提供了一种新的思路。本研究通过模拟比较了 GDINA、Mixed、DINA、DINO、ACDM 和 CRUM 在诊断测验中的效果，结果发现，真模型是简化模型时，简化模型的效果是最好的，但当真模型是 GDINA 或者 Mixed 时，选用 Mixed 方法选择的模型具有更好的效果。对 Tatsuoka 分数减法实测数据从 3 个方面比较了几种单模型和混合方法的整体表现：模型资料相对拟合指标（偏差、AIC、BIC），题目与模型的绝对拟合指标（Chen, de la Torre, & Zhang, 2013），测验信度指标（Cui, Gierl, & Chang , 2012），结果发现混合方法的综合表现是最好的。从理论和实际应用的角度混合方法都具有一定的参考价值。尽管如此，但并不意味着混合方法在任何情况下都是最优的选择，混合方法是一种基于统计指标检验的数据驱动的方法，数据驱动方法的优势，同时也是劣势，即普适性较差，混合方法是基于 Wald 检验为每一个题目选择模型，因而存在一定的缺陷，如样本很少的情况下，Wald

检验容易出现较大误差（可能为每个题目选出的模型并不拟合），在样本较少或者测验质量并不理想的情况下，此时，混合方法选出的模型可以结合其他拟合指标使用，如首先从测验整体拟合的角度（计算 AIC、BIC 指标）比较混合方法选出来的模型和其他单模型的效果，从而验证混合方法选择的模型是不是相对拟合较好的，如果是的话，进一步从题目拟合的角度对混合方法选择的模型展开验证，即比较单个题目和模型的绝对拟合度，如果发现有题目和混合方法选出的模型并不拟合，则可能需要进一步分析不拟合的原因，或试着通过其他方法（如比较其余模型在这些题目的绝对拟合指标），为这些题目找到一个更优的模型。

本研究所得的实验结论建立在 Q 矩阵正确的前提下，Q 矩阵在诊断测验中扮演着非常重要的角色，Q 矩阵是否正确直接影响到诊断评估的准确性，Q 矩阵错误标定对混合方法的影响需要更进一步的研究。另外，考虑到混合方法是基于 Wald 统计检验的方法，作为一种数据驱动方法，其效果会受到样本大小的影响，本研究最小样本是 500，样本更少的情况下，混合方法是否也有同样的效果有待进一步研究。本研究假设属性之间并不存在层级关系以及被试的知识状态服从均匀分布，在实际测验中属性之间的结构和被试知识状态的分布会更复杂，所以本研究的结论还有待进一步的验证。

参考文献

高旭亮，涂冬波 . (2017). 参数化认知诊断模型：心理计量特征，比较及其转换 . *江西师范大学学报：哲学社会科学版，50(1),* 88–104.

涂冬波，蔡艳，戴海琦 . (2013). 几种常用非补偿型认知诊断模型的比较与选用：基于属性层级关系的考量 . *心理学报，45(2),* 243–252.

Chen, J., 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., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49(1),* 19–38.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76(2),* 179–199.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item–level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50(4),* 355–373.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26(4),* 301–321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Doctoral dissertation, University of Illinois at Urbana–Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log–linear models with latent variables. *Psychometrika, 74(2),* 191–210.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to Language Testing assessment. *Language Testing, 26(1),* 31–73

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32(2),* 227–258.

Lei, P. W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q–matrices. *Applied Psychological Measurement, 40(6),* 405–417.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33(3),* 391–409.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69(3),* 253–275.

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

Rojas, G., de la Torre, J., & Olea, J. (2012l). Choosing between general and specific cognitive diagnosis models when the sample size is small. *Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada.*

Tatsuoka, K. K. (1990). Toward an integration of item–response theory and cognitive error diagnosis. *Diagnostic monitoring of skill and knowledge acquisition*, 453–488.

Templin, J. L., & Henson, R. A. (2006).Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11(3),* 287–305.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61,* 287–307.

Zhang, J. (2013). *Relationships between missing responses and skill mastery profiles of cognitive diagnostic assessmen*t. Doctoral dissertation, University of Toronto.

Comparison of CDM and its Selection: A Saturated Model, A Simple Model or A Mixed Method

Gao Xuliang, Wang Daxun, Cai Yan, Tu Dongbo

(Research Center of Psychological Health Education, School of Psychology, Jiangxi Normal University, Nanchang, 330022)

Abstract Recent advances in a category of analytic methods collectively referred to as cognitive diagnostic models (CDM) show great promise. A large number of CDM have been proposed, The deterministic inputs, noisy, ‘‘and’’ gate (DINA) model, an example of a conjunctive model, assigns the highest probability of answering correctly to examinees that possess all of the required attributes. Disjunctive models, however, assume that lacking a particular attribute can be off-set by possessing another. For example, the deterministic inputs, noisy, ‘‘or’’ gate (DINO) model assigns the highest probability of answering correctly to examinees with at least one of the required attributes. Examples of other specific, interpretable CDM are the reduced reparametrized unified model (CRUM; Hartz, 2002), the additive CDM (ACDM). Apart from these specific CDM, general or saturated CDM subsuming many widely used specific CDM have also been developed, including the generalized DINA (GDINA) model, the general diagnostic model (GDM), and the log-linear CDM (LCDM). Although general CDM provide better model-data fit, reduced CDM have more straightforward interpretations, are more stable, and can provide more accurate classifications when used correctly.

Although a multitude of CDM are available, it is not clear how the most appropriate model for a specific test can be identified because the cognitive processes in answering items may be complicated. An important decision that researchers make is that of choosing either a CDM that allows for compensatory relationships among skills or one that allows for non-compensatory relationships among skills. With a compensatory model, a high level of competence on one skill can compensate for a low level of competence on another skill in performing a task. Specifically, a general model (i.e., GDINA model) can be tested statistically against the fits of some of the specific CDM it subsumes using the Wald test. The Wald test was originally proposed by de la Torre (2011) for comparing general and specific models at the item level (i.e., one item at a time) thereby creating the possibility of using multiple CDM within the same test which means each item selects a different model. In order to compare the mixed method and other model performance in the paper and pencil test, using a complex simulation study we investigated parameter recovery, classification accuracy, and performance of item-fit statistics for correct and misspecified diagnostic classification models within a GDINA framework. The basic manipulated test design factors included the number of respondents, item quality generating model, fitted model and Q-matrix. The three sample sizes were N = 500 and 1,000, item quality were high and low, generating model and fitted model were GDINA, Mixed, DINA, DINO, ACDM and CRUM, Q-matrix included simple Q-matrix and complex Q-matrix. The study found that overall under all experimental conditions, the Mixed CDM had the best performance. Simply take into account classification accuracy rate, Mixed in low quality advantage is more obvious in the tests, when item quality is high, Mixed and GDINA performance is almost identical, but under all experimental conditions, mixed was better than GDINA in information-based fit indexes AIC and item parameter recovery.

Key words CDM , GDINA, Wald test, Reduced CDM