《心理学报》审稿意见与作者回应

题目：认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标
作者：郭磊　郑蝉金　边玉芳　宋乃庆　夏凌翔

第一轮

审稿人 1 意见：

《认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标》评审意见：本文选题具有意义，创新性较高，可读性强。但是读者还有以下问题和建议：

意见 1：3.5 最后一段话提到“公式 12 可能为负，导致 MIj 排序错误”，是什么意思，请给与解释。另外，我认为公式 12 都取负值，请问什么时候为正？

回应：感谢审稿人提出的意见。首先，在修改本文过程中，我们发现 Wang（2013）论文中的公式（13）和（14）存在排版/印刷错误，Wang 原文中的公式（13）和（14）中的 $\frac{\log p\left(X_t=x\mid\boldsymbol{\alpha}_l\right)}{h_2}$ 部分实际上应该是 $\log\frac{p\left(X_t=x\mid\boldsymbol{\alpha}_l\right)}{h_2}$，在推导过程中，h2 不可能推至 log 运算之外，这一点我们和 Wang 本人确认过，发表的文章中确实存在该印刷错误。其次，为了回答审稿人的问题，我们需要理解 Wang 是如何化简公式的。现在从原始公式开始：

$$MI_j=\sum_{x=0}^{1}p\left(X_t=x\mid\mathbf{x}_{t-1}\right)\left[\sum_{l=1}^{2^K}\pi\left(\boldsymbol{\alpha}_l\mid\mathbf{x}_{t-1},X_t=x\right)\log\left(\frac{\pi\left(\boldsymbol{\alpha}_l\mid\mathbf{x}_{t-1},X_t=x\right)}{\pi\left(\boldsymbol{\alpha}_l\mid\mathbf{x}_{t-1}\right)}\right)\right] \quad (100)$$

其中，$p\left(X_t=x\mid\mathbf{x_{t\text{-}1}}\right)=\sum_{l=1}^{2^K}p(X_t=x\mid\boldsymbol{\alpha}_l)\pi\left(\boldsymbol{\alpha}_l\mid\mathbf{x}_{t-1}\right)=\frac{\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1},X_t=x\mid\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1}\mid\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}$，

$$\pi\left(\boldsymbol{\alpha}_l\mid\mathbf{x}_{t-1},X_t=x\right)=\frac{p(\mathbf{x}_{t-1},X_t=x\mid\boldsymbol{\alpha}_l)\,p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1},X_t=x\mid\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}$$

并且规定：$h_1=\sum_{l=1}^{2^K}p\left(\mathbf{x}_{t-1}\mid\boldsymbol{\alpha}_l\right)p\left(\boldsymbol{\alpha}_l\right)$，$h_2=\sum_{l=1}^{2^K}p\left(\mathbf{x}_{t-1},X_t=x\mid\boldsymbol{\alpha}_l\right)p\left(\boldsymbol{\alpha}_l\right)$。那么，原始公式（100）可以改写为：

$$MI_j = \sum_{x=0}^{1} \frac{h_2}{h_1} \left[\sum_{l=1}^{2^K} \frac{p(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l)}{h_2} \log \frac{p(X_t = x \mid \boldsymbol{\alpha}_l) h_1}{h_2} \right]$$
$$= \frac{1}{h_1} \left[\sum_{x=0}^{1} \sum_{l=1}^{2^K} p(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l) \log \frac{p(X_t = x \mid \boldsymbol{\alpha}_l)}{h_2} + \log h_1 \sum_{x=0}^{1} \sum_{l=1}^{2^K} p(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l) \right] \quad (101)$$

在公式（101）中，红框内为修改正确后的表达式（注意：h2 在 log 运算里面），蓝框中有：$\sum_{x=0}^{1} \sum_{l=1}^{2^K} p(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l) = \sum_{l=1}^{2^K} p\left(\mathbf{x}_{t-1} \mid \boldsymbol{\alpha}_l\right) p\left(\boldsymbol{\alpha}_l\right) = h_1$，因此，公式（101）可以进一步改写为：

$$MI_j = \sum_{x=0}^{1} \frac{h_2}{h_1} \left[\sum_{l=1}^{2^K} \frac{p(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l)}{h_2} \log \frac{p(X_t = x \mid \boldsymbol{\alpha}_l) h_1}{h_2} \right]$$
$$= \frac{1}{h_1} \left[\sum_{x=0}^{1} \sum_{l=1}^{2^K} p(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l) \log \frac{p(X_t = x \mid \boldsymbol{\alpha}_l)}{h_2} + h_1 \cdot \log h_1 \right] \quad (102)$$

因为 h1 是一个常数，不会受到当前所选第 t 题的影响，因此，Wang 将公式（102）中蓝框内的表达式删除掉，就导出了最终化简版的 MI 指标公式：

$$MI_j = \frac{1}{h_1} \sum_{x=0}^{1} \left[\sum_{l=1}^{2^K} p\left(\mathbf{x}_{t-1}, X_t = x \mid \boldsymbol{\alpha}_l\right) p\left(\boldsymbol{\alpha}_l\right) \log \frac{p\left(X_t = x \mid \boldsymbol{\alpha}_l\right)}{h_2} \right] \quad (103)$$

化简版的 MI 值是正是负，取决于 $\frac{p(X_t = x \mid \boldsymbol{\alpha}_l)}{h_2}$ 的比值是否大于 1。若该比值小于 1，则 $\log \frac{p(X_t = x \mid \boldsymbol{\alpha}_l)}{h_2}$ 为负值；若比值大于 1，则 $\log \frac{p(X_t = x \mid \boldsymbol{\alpha}_l)}{h_2}$ 为正值。

产生“错误排序”的原因是：在化简过程中，丢掉了蓝框内的表达式 $h_1 \cdot \log h_1$，若未加入曝光因子或非统计限制，单纯看 MI 指标的大小，并不会产生乱序。但在乘以曝光因子之后，丢掉的部分将不再是 $h_1 \cdot \log h_1$，而是 $h_1 \cdot \log h_1$ 再乘以曝光因子的大小，记作 $f_j \cdot h_1 \cdot \log h_1$，$f_j$ 为题目 j 当前的曝光因子大小。然而，每个题目的曝光因子在连续施测的 CAT 测验模式下是动态变化的，即大小是不一样的，所以丢掉部分的大小也不一样，此时，可以理解为量尺发生了变化，这就会导致题目产生乱序。“乱序”概念是相对于使用原始公式计算后的正常排序而言的。本研究在进行模拟实验时，加入了项目曝光控制，因此，如果使用化简版公式，将导致乱序现象的出现。因此，在仔细思考后，我们决定改用原始公式进行 MI 值的计算，

这样即使乘以了曝光因子或其他非统计限制因子，也绝对不会出现乱序现象。相应改动请见修改稿 3.5 和 4.2 部分。

意见 2：4.2 研究设计第三段提到的 PR 方法，是本文结合了 RP 方法与新指标选题吗？RP 本来就有一个选题指标，请问你是怎么结合使用的？另，此段末尾，MI 值应该加上|max（MIj）|还是加上|min（MIj）|?

回应：感谢审稿人的意见，本文正是像您说的是“结合了 RP 方法与新指标选题”，我们在原文中没有表达清楚，现已将该部分内容在文中进行了修改，请见修改稿 4.2 部分。另外，原稿中的|min（MIj）|是 MATLAB 中求最小值的内置函数，其实质就是加上最小的负值的绝对值。但由于修改稿中，我们不再使用化简版的 MI 公式而改用原始公式（理由请见对您的第一个回答），因此就不再涉及 MI 需要加上一个常数的问题了。相应的 MIM 结果也全部重新进行了实验，请见表 2、表 3、表 4 和表 5。

意见 3：4.3 是不是应该补充对测验重叠率和卡方的介绍？

回应：感谢审稿人的意见，我们已在修改稿 4.3 部分增加了对测验重叠率以及卡方的介绍。

意见 4：第五部分，也采用了 simple 方法控制项目曝光率，请问你的曝光率结果呢？

回应：感谢审稿人提出的意见。我们没有将变长 CDCAT 的曝光率结果写进原文中，是基于两个原因：①在比较不同的选题策略质量差异时（控制其他条件均相同），若选用定长终止规则，那么判准率高的选题策略较好；若选用变长终止规则，即在固定终止精度时，主要看平均用题量，即平均用题量少的选题策略较好。因此，本文只将最主要的考察/比较指标呈现出来（请见正文表 4 和表 5）。②当加入曝光控制技术后，我们可以预期曝光率是能够控制在 r_{max} 以下的，这也正是加入曝光控制的目的。另外，为了让审稿人及读者能了解到加入 simple 法后的曝光率，我们将该结果呈现在该回答中，请见表 6。如表 6 所示，所有实验条件下的最大曝光率均小于本文预设的曝光率上限 0.2。

表 6　simple 法曝光控制下的曝光率

结构和质量	P_{1st}	PWKL	GIDPWKL	AIDPWKL	CIDPWKL	KLEDPWKL	MIM
		Max(r)	Max(r)	Max(r)	Max(r)	Max(r)	Max(r)
S-H	0.7	0.080	0.088	0.087	0.086	0.099	0.073
	0.8	0.085	0.093	0.091	0.090	0.104	0.090
	0.9	0.087	0.100	0.098	0.095	0.109	0.095
S-L	0.7	0.098	0.105	0.104	0.110	0.116	0.096
	0.8	0.106	0.109	0.113	0.114	0.126	0.107
	0.9	0.111	0.116	0.120	0.119	0.132	0.128
C-H	0.7	0.095	0.135	0.129	0.106	0.127	0.103
	0.8	0.096	0.138	0.130	0.107	0.128	0.114
	0.9	0.098	0.141	0.139	0.110	0.131	0.125
C-L	0.7	0.120	0.156	0.145	0.132	0.143	0.128
	0.8	0.125	0.157	0.149	0.133	0.145	0.131
	0.9	0.125	0.161	0.150	0.135	0.148	0.136

注：Max(r)表示最大曝光率

意见 5：读者感觉 4.4 的结果为没有用到曝光控制方法的结果，不知道您的结果是否正确？

回应：本研究的结果是正确的。为了节省文章篇幅，我们没有将未加入曝光控制的结果放入正文中，该部分结果在该回答中呈现，请见下表 7 和表 8。通过与正文中的表 2 和表 3 结果比较，在相同实验条件下未加入曝光控制的判准率（AACCR、PCCR）均有不同程度的上升，而题库使用情况会变差，具体表现为测验重叠率、题库中未使用的题目数量，以及卡方值均会增大。该情况正是 CAT 中常见的估计精度和题库使用情况的权衡（*trade-off*）问题。

表 7 简单结构下不同选题策略的判准率及题库使用情况（未加入曝光控制的结果）

测验长度	题目质量	选题方法	AACCR	PCCR	T	NU	Chi
5	高	PWKL	0.864	0.421	0.588	763	465.70
		GIDPWKL	0.891	0.479	0.882	783	700.65
		AIDPWKL	0.877	0.465	0.795	777	617.60
		CIDPWKL	0.870	0.462	0.753	769	583.15
		KLEDPWKL	0.868	0.460	0.736	766	554.49
		MIM	0.876	0.465	0.788	771	609.12
	低	PWKL	0.788	0.345	0.582	764	461.24
		GIDPWKL	0.816	0.379	0.782	783	600.64
		AIDPWKL	0.797	0.368	0.678	773	525.59
		CIDPWKL	0.791	0.360	0.639	769	501.79
		KLEDPWKL	0.788	0.359	0.631	766	492.70
		MIM	0.795	0.362	0.656	770	516.00
10	高	PWKL	0.966	0.828	0.602	716	452.02
		GIDPWKL	0.980	0.862	0.676	742	522.10
		AIDPWKL	0.974	0.851	0.628	715	476.24
		CIDPWKL	0.979	0.859	0.639	730	501.70
		KLEDPWKL	0.975	0.853	0.637	727	496.41
		MIM	0.976	0.857	0.637	728	498.76
	低	PWKL	0.871	0.552	0.580	710	454.62
		GIDPWKL	0.892	0.579	0.676	741	522.24
		AIDPWKL	0.887	0.572	0.669	736	504.43
		CIDPWKL	0.880	0.566	0.594	718	473.38
		KLEDPWKL	0.883	0.571	0.663	733	497.59
		MIM	0.889	0.574	0.670	734	518.70

注：T 为测验重叠率，NU 为题库中未使用的题目数量，Chi 为卡方值

表 8 复杂结构下不同选题策略的判准率及题库使用情况（未加入曝光控制的结果）

测验长度	题目质量	选题方法	AACCR	PCCR	T	NU	Chi
5	高	PWKL	0.827	0.374	0.522	737	397.76
		GIDPWKL	0.866	0.438	0.618	787	469.81
		AIDPWKL	0.858	0.427	0.599	785	450.20
		CIDPWKL	0.839	0.414	0.564	754	439.53
		KLEDPWKL	0.833	0.406	0.551	747	425.80
		MIM	0.853	0.423	0.573	766	443.31
	低	PWKL	0.744	0.295	0.549	722	428.73
		GIDPWKL	0.780	0.357	0.633	785	479.62
		AIDPWKL	0.777	0.353	0.592	780	471.72
		CIDPWKL	0.756	0.342	0.560	734	436.54

		KLEDPWKL	0.759	0.345	0.569	752	443.42
		MIM	0.751	0.336	0.545	713	416.31
10	高	PWKL	0.955	0.806	0.484	655	377.33
		GIDPWKL	0.976	0.851	0.624	749	465.70
		AIDPWKL	0.969	0.845	0.567	722	431.76
		CIDPWKL	0.964	0.838	0.540	696	407.06
		KLEDPWKL	0.964	0.839	0.545	708	415.95
		MIM	0.967	0.842	0.552	717	426.40
	低	PWKL	0.856	0.528	0.527	640	411.76
		GIDPWKL	0.873	0.560	0.658	733	489.63
		AIDPWKL	0.871	0.555	0.625	702	451.41
		CIDPWKL	0.862	0.545	0.609	676	430.74
		KLEDPWKL	0.867	0.551	0.611	681	434.80
		MIM	0.870	0.553	0.623	693	448.77

审稿人 2 意见：

本文存在如下问题

意见 1：文章的创新性有待提高。在选题策略中考虑项目区分度指标，类似研究也有（见如汪文义，丁树良和宋丽红，2014），在汪文义等人（2014）研究中就考虑了选题策略项目区分度（1-s-g）加权问题。

回应：感谢审稿人提出的意见。我们在修改稿中将汪文义等人（2014）的方法进行了介绍，并纳入到了本研究的比较中，请见本文引言及方法等相关部分。

本文理论上的创新性正如“论文自检报告”中所述：当前在 CD-CAT 中使用的选题策略属于单源(single-source)指标，而将项目区分度信息融合到单源指标中的多源(multiple-source)选题指标正是本文的创新之处。具体而言，首先，本文使用的（1-s-g）加权思想是来源于 Rupp 等（2012）书中第 13 章里介绍的基于 CTT 的项目区分度指标，在加权的形式上与汪文义等人（2014）提出的 $(1-\mathrm{s}-\mathrm{g})(\log\frac{1-s}{g}+\log\frac{1-g}{s})$ 形式略有不同，并且基于 Rupp 等提出的基于 CTT 的项目区分度思想上，不仅可以将该加权方法用于 DINA 模型，还可以拓广至其他模型，例如本文第 6 部分提到的融合模型，其适用性更广，而汪文义等提出的权重 $(1-\mathrm{s}-\mathrm{g})(\log\frac{1-s}{g}+\log\frac{1-g}{s})$ 只适用于 DINA 模型。其次，汪文义等人（2014）提出的加权方法仅是基于 CTT 思想的项目区分度加权方法，并没有考虑基于 KL 信息量的全局项目区分度加权法和基于 KL 信息量的属性层面项目区分度加权法。根据本文的研究结果可以看出，使用基于 KL 信息量的项目区分度加权方法在定长测验情境下能够得到更高的判准率，在变长测验情景下能够缩短测验题目，这都是本文的重要贡献。第三，汪文义等人（2014）提出的加权方法是基于 KL 指标的加权，而非 PWKL 指标的加权。第四，Wang(2013)提出的互信息选题策略（MIM）计算公式相当复杂，不利于读者理解，因此，寻求一个更易理解、且效果更好的方法是非常有必要的。第五，Wang(2013)虽提出了 MIM，并通过模拟研究得到了 MIM 在较短测验中表现要优于 PWKL 的结果，但目前缺乏对本研究提出的新方法、MIM 以及汪文义等人（2014）方法的全面比较。

本文实践上的创新性体现在：在实际教学过程中，教师往往希望通过较短的几道题目就

能获得对学生知识状态的了解，然后有针对性地对课堂教学内容进行调整和把控，因此开发一个高效的选题方法十分重要。这里的高效指在较短测验上获得更高的估计精度。因为任何选题方法，当测验题目不断增加时，其判准率都会逐渐提升，选用较长测验进行比较研究，就不能凸显不同选题方法之间的差异。因此，本研究聚焦于较短测验，在较短测验上有较高判准率的方法是性能更好的方法。

意见 2：PWKL 指标是一个综合性的信息量指标，它不仅包含了项目区分度信息，还考虑了被试当前知识状态与项目测量模型的匹配问题，这一点文章并未认识到。而文章在 PWKL 中再对项目区分度进行加权，这无非导致测验过度使用题目质量好的项目，这个不符合 CAT 设计思想。

回应：感谢审稿人的意见。通过对指标的分析，我们可以知道，KL 指标中包含了题目参数以及被试当前知识状态与其他知识状态之间的 KL 距离的信息，而 PWKL 在 KL 的基础上增加了后验概率的信息。在 IRT 中，项目参数 a 表示项目区分度，但在认知诊断中，题目参数本身并不直接等价于项目区分度，具体来说，基于 CTT 思想的项目区分度实际上是题目参数的函数，而基于 KL 信息量的全局项目区分度（请见本文公式 6）和基于 KL 信息量的属性层面的项目区分度（请见本文公式 8）更是和 KL（PWKL）中的题目参数有本质上的不同。例如，基于 KL 信息量的全局项目区分度 $C_j = \frac{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} \cdot D_{j,uv}}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}}$，其中不仅包含了两两 KS 之间的 KL 信息量（提供的信息要远远大于只考虑被试当前知识状态与其余知识状态的 KL 信息量），还将两两 KS 之间的汉明距离也纳入了考量，这些都与 PWKL 存在着本质的区别。

正是因为项目区分度和 PWKL 指标是从两个不同的角度刻画题目的“区分能力”，因此将两者结合是合理的。事实上，我们从程序中截取了中间一段结果，发现 $PWKL_j$ 值高的题目，其全局项目区分度 C_j 值不一定是高的，这也避免了审稿人指出“导致测验过度使用题目质量好的项目”的问题。另外，在本文的模拟研究中，我们加入了对题目曝光的控制，使得质量好的题目不会过度曝光。

意见 3：更为重要的是，文章将区分度指标加入到 PWKL 选题策略中后，使得测验的曝光率更大，题库的安全性更差，而诊断正确率平均提高的幅度却还不到 4%（最大的也就 5.8%）。这说明新选题策略是以牺牲题库安全性为代价，且属性判准率提高的也非常少。

回应：感谢审稿人提出的意见。在 CAT 中，估计精度和题库使用情况之间存在着权衡（*trade-off*）问题。若研究目的在于提升估计精度，那题库使用情况势必会受到影响。若更加注重题库的使用，估计精度又会受到影响，这在所有的 CAT 研究中都是普遍存在的现象。在这两者之间如何抉择，需要根据研究目的和实际使用情景来定。本文的研究目的及重点是在较短测验时，如何能够快速地提升对被试 KS 的估计精度，题库的使用会受到影响是可以预期的。根据研究结果可以看出，在提高判准率的同时，测验重叠率和卡方值比起 PWKL 的结果上升得并不多。另外，本文提出的方法是用于实际课堂教学中让教师快速地了解学生当前的知识掌握情况，因此，比起估计精度，题库的使用情况并不是本研究关注的重点，而且本研究在进行实验时考虑了项目曝光控制技术，得到的题库使用结果是在预期控制范围内的。

本研究的实验条件是聚焦于较短测验长度，分别是 5 题和 10 题。根据研究结果可以看出，测验长度越短，新方法的有效性越好。本研究只用了 5 道题目就能提升 4%~5%（最高 5.8%）的判准率，并且结果比 MIM 方法还要好（MIM 是目前认知诊断研究中，在较短测

验中最好的方法，本文研究是和当前最好的方法作比较），足以说明在测验前期，加入项目区分度信息将能够快速地提升对被试 KS 的估计精度，在实际教学中具有较高的实用价值。可以预期的一点是，若 CDCAT 题库中的项目区分度更好，新方法在短测验中的优势将更明显。

意见 4：本文可读性不强，文章逻辑不够清晰。
回应：感谢审稿人提出的意见。我们对正文的逻辑以及文字等方面进行了修改，引言部分的最后一段陈述了文章的写作逻辑。

审稿人 3 意见：
本文将项目区分度加入选题策略中，克服了以往仅仅根据难度选题的局限。模拟研究方法设计上也较为严谨，得出了一些有意义的结论。但也有些问题需要考虑（建议在讨论部分加以说明）：
意见 1：难度与区分度是动态的关系。即在不同的能力水平上，一个题目的区分度也应会有所变化。因此选题兼顾二者一般很难实现。
回应：感谢审稿人提出的问题。您说的情况在项目反应理论（Item Response Theory，IRT）确实存在。在实际题库中，区分度 a 与难度 b 通产都是正相关的(Lord, F. M. (1975). The 'ability' scale in item characteristic curve theory. Psychometrika, 40(2), 205–217.)。但在认知诊断中，一般不提项目的难度参数（至少基于当前认知诊断的相关研究，并未定义出难度参数），因此，难度与区分度之间并没有动态关系，也就不存在选题需要兼顾二者的做法。从认知诊断中对于项目区分度的定义可以看出，区分度是一个项目自身的参数，当考察的属性个数在某次测验中确定之后，该参数并不会受到被试知识状态的影响。例如，基于 KL 信息量的全局项目区分度为：$C_j = \frac{\sum_{u\neq v} h(\boldsymbol{\alpha}_{\mathbf{u}}, \boldsymbol{\alpha}_{\mathbf{v}})^{-1} \cdot D_{j,uv}}{\sum_{u\neq v} h(\boldsymbol{\alpha}_{\mathbf{u}}, \boldsymbol{\alpha}_{\mathbf{v}})^{-1}}$，该计算公式并不依赖于被试当前估计的知识状态，在被试进行测验之前，就可以将所有项目的区分度求出来了。若我们没有很好地理解您该问题的具体意思，请您进一步详细说明。

意见 2：如果研究者事先就选择了高区分度的项目进入题库，则根据区分度选题的做法，其实际意义就打了折扣。
回应：感谢审稿人提出的问题。本研究的出发点是探讨如何在较短测验中快速提升被试知识状态的估计精度。新的选题方法并不仅是基于项目区分度进行选择，同时还考虑了 PWKL 信息量，选题是两者综合考量的结果。正如我们回答第二位审稿人的第 2 个问题所述，$PWKL_j$ 值高的题目，其全局项目区分度 C_j 值不一定是高的，因此，我们其实无法做到"事先就选择高区分度的项目"。

意见 3：所以请重新思考本文的研究立意。
回应：感谢审稿人提出的问题。我们已在引言部分突出了本研究的立意，即如何能在较短测验中准确地估计学生的知识状态，为形成性评估提供实际帮助，这就需要高效的选题方法作为支持。另外，我们也已根据三位审稿人的意见，对论文进行了大量修改，使本文质量有了进一步提高。

意见 4：写作格式上也要再规范些。

回应：感谢审稿人提出的意见。我们已在修改稿中做了大量修改。

最后，再次感谢三位审稿人的宝贵意见，使得修改稿较原稿在整体质量上得到了较大提升。

第二轮

审稿人 1 意见：

《认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标》

本文阅读流畅，结构符合逻辑，实验设计方法合理，文中公式推导正确，结果可信。

意见 1：但是作者有待思考，本文的创新性和意义如何？另外，论文还有几处有待调整：

回应：感谢审稿人对本文写作、设计、数理公式以及结果的肯定。本文的创新和意义正如自检报告中所述：（1）当前在CD-CAT中广泛运用的选题策略（主要指PWKL）属于单源(single-source)指标，仅利用了个体层面信息对KL信息量进行加权，并未考虑到题目质量会给选题过程带来的影响。其实，在PWKL基础上再进行加权的做法和思想在Cheng（2009）[Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing. Psychometrika, 74, 619-632.]一文中就已经提出，她提出了后验加权和距离加权的混合KL信息量方法（HKL），HKL可以看作个体层面的双加权指标，因为距离加权部分仍然利用的是个体层面的信息。但根据实验结果，HKL和PWKL的表现几乎一致，说明只从个体层面进行加权（不论几次）的做法并不能再提供额外信息了。而本文突破性地提出了结合个体层面和题目层面的双加权指标，在PWKL的基础上，同时考察了题目质量对选题过程产生的影响，这是创新点之一。（2）本文将多种项目区分度信息融入PWKL指标，和Wang（2013）提出的互信息法（互信息法是目前认知诊断研究中，在较短测验中最好的方法，本文研究是和当前最好的方法作比较）在定长和变长两种不同的测验情景中进行对比研究，并考察了多种实验因素，实验设计较充分，这也是创新点之一。

本研究的重要意义是：在实际教学过程中，教师往往希望通过较短的几道题目就能获得对学生知识状态的了解（快速、精准），从而有针对性地对课堂教学内容进行调整和把控，因此开发一个高效的选题方法十分重要。这里的高效指在较短测验上获得更高的估计精度。因为任何选题方法，当测验题目不断增加时，其判准率都会逐渐提升，选用较长测验进行比较研究，就不能凸显不同选题方法之间的差异。因此，本研究聚焦于较短测验，在较短测验上有较高判准率的方法是性能更好的方法。

相关内容还可以参见第一轮对第二位审稿人的第一条意见的回答。

意见 2：结果描述显得太复杂，需要精简。

回应：感谢审稿人的意见，我们已对结果描述部分进行了精简。4.4.部分比上次修改稿精简了 402 字，5.4 部分比上次修改稿精简了 57 字。

意见 3：3.2 部分第一段第四行叙述属性区分度的表述不够清晰，需要调整。

回应：感谢审稿人的意见，我们已增加了对 CDA 中项目区分度的解释，以帮助读者进行理解。

具体内容为：但在 CDA 中，测评的结果是以多维离散潜在变量呈现，不像 CTT 那样能够基于总分找到高分组和低分组。因此，Rupp 等(2012)仿照 CTT 的思想，定义了 CDA 中的项目区分度：$d_j = p_{\alpha_h} - p_{\alpha_l}$。其中，$p_{\alpha_h}$ 表示掌握题目 j 考察的较多属性的正确作答概率，p_{α_l} 表示掌握题目 j 考察的较少属性的正确作答概率。其含义是：“该项目区别掌握‘较

多’属性被试和掌握‘较少’属性被试的能力”。

意见 4：3.3 部分公式 6 中的 D 是个矩阵，那么公式 7 对每个项目而言将是一个矩阵，无从比较。请检查是表述错误，还是公式 7 比较的是什么？请说明。
回应：感谢审稿人的细致审查，这确实是表述错误，我们已经在原文相应位置进行了修改。

意见 5：5.3 评价指标应该讲清楚是什么的平均数，标准差等等。
回应：感谢审稿人的意见，我们已增加了对平均数，标准差等概念的描述。该部分阐述同样可以参见 5.1 部分。

审稿人 2 意见：

经过作者的修改，本文的科学性和严谨性得到提高。特别是作者补充了对汪文义等人提出方法的模拟，使文章的内容更为丰富，参考价值也更大。文章的研究方法正确、思路清楚，评判指标选择合理，其结果对发展 CD-CAT 的选题策略有重要贡献。但文章的结构和写作方面还有可以改进的地方。
意见 1：文章前 3 部分内容和体量上不平衡，引言中对一些选择策略的介绍和述评比较详细，但在相关选题策略介绍部分，再次提及某些策略时只是列出和描述了公式，可以适当增加述评。另外，第 2 部分“DINA 模型简介”只有一段话，是否有单独成节的必要，是否可以考虑与第 3 部分放在一起。
回应：感谢审稿人的意见。引言部分之所以进行详细介绍和述评，是按照国际上本领域的写作要求展开的，以期较详尽的将当前该研究问题的背景和研究意义介绍清楚。选题策略部分，即第三部分内容共包含了 6 种选题策略的介绍，如果展开介绍，那么篇幅将会增加不少。我们在介绍每种方法时，都有加入对该方法的简要说明和介绍（已标记为紫色），不知您是否满意。

DINA 模型简介这部分内容，一开始我们是想和第三部分放在一起，但考虑到“模型”和“方法”毕竟是两个不同的内容，而且我们参考了 APM 和 EPM 杂志上的写作风格，例如 Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. Educational and Psychological Measurement. 73(6),1017-1035.，以及 Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. Applied Psychological Measurement, 37(7), 563-582 等人的文章，均是单独的介绍认知诊断模型，因此，我们最后仍决定将 DINA 模型进行单独介绍。

意见 2：文中的一些措辞过于绝对。如引言中“查阅国内外相关文献，只有汪文义等(2014)基于 CTT 的思想将项目区分度信息纳入选题策略中进行了研究”，这样判定为“只有”的话是否有充分的根据。最后的结论与讨论部分中“与传统单维 IRT 模型相比，认知诊断模型能够提供更加丰富的诊断信息，用于随后的补救教学。”将 CD 测验的作用限定在了补救教学上，也不完整。
回应：感谢审稿人的意见。按照国外高水平论文中看到的相关陈述方式，例如 To our best knowledge（Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*, 28-54.），我们将第一句修改为：“查阅国内外相关文献，将区分度信息纳入 CD-CAT 选题过程的研究并不多，据我们所知，汪文义等(2014)基于 CTT 的思想将项目区分度信息纳入选

题策略中进行了研究”。

结论与讨论部分的第一句话，我们修改为：与传统单维 IRT 模型相比，CDA 除了能够提供单维的能力估计值以外，还能获得被试在知识点上的掌握情况，所提供的测评信息更加丰富，为教育教学工作提供有针对性的指导，用于补救教学中，促进学生的个性化发展。

意见 3：作者修正的 4 个指标中，是否都会优先选用区分度更高的题目？文章中作者也提到，实际上也有一种思想，a-分层选题法，先不选用区分度高的题目。关于这方面，是否可以增加必要的说明和自己的思考。

回应：感谢审稿人专业的提问。本研究修正的 4 个指标（GIDPWKL、AIDPWKL、CIDPWKL 和 KLEDPWKL）并不会优先选用区分度更高的题目。通过公式（5）、（7）、（9）、（15）可以看出，优先选择的题目是“乘积”值最大的题目，“乘积”是由两个因素决定的：项目区分度和 PWKL 值。并且，想要优先选用区分度更高的题目也并非本研究的初衷。事实上，我们从程序中截取了中间一段结果，发现 PWKL 值高的题目，其区分度不一定是大的，同样，区分度大的题目，其 PWKL 值也不一定是高的。

本文在引言中介绍 a 分层选题法的目的是强调项目区分度在选题过程中的重要性，但该方法的思想和本文所使用的方法的思想是不一样的，很多研究者也研究了如何在认知诊断中进行分层选题，请参见毛秀珍，辛涛(2013)在心理学报发表的文章《认知诊断 CAT 中项目曝光控制方法的比较》，以及汪文义，丁树良，宋丽红（2014）在心理科学发表的文章《兼顾测验效率和题库使用率的 CD－CAT 选题策略》。

意见 4：文章最终的总结论实际为“多源指标是更加有效的选题策略，在实际应用中应该首选测验效率最高的 GIDPWKL 方法。”其实，定长实验中，GIDPWKL 方法的判准率也是最高的，是否在最终结论中也强调方法的准确性。

回应：感谢审稿人的建议，这更加能突出 GIDPWKL 方法的优越性。我们已在论文中增加了该内容。

意见 5：最后，建议作者在写作中可以进一步提高语言的多样性。不宜在多个并列段落中完全使用同样的句式，只是更换其中的个别关键词，这样容易让人产生阅读疲劳。特别是本文中一些英文简写又很相近，还容易让人发生阅读错误。

回应：感谢审稿人的意见，我们已经通读过全文，同时邀请了三位相关专业的硕士研究生和博士研究生对论文进行了精读，将容易犯错误的地方进行了严格检查。

审稿人 3 意见：

《认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标》一文把项目区分度指标做为权重，与已有选题策略相结合。这种从项目参数出发，来重构选题策略是一种新的思路。通读全文，还有以下不清楚之处，请作者予以解释。

意见 1：为什么要将项目区分度作为一个考量因素，仅仅是因为在传统 CAT 当中有按 a 分层，故而在 CD-CAT 中也沿用这个吗？其次，为何将其做为权重进行考量等，请作者对此进行详细说明。

回应：感谢审稿人的提问。我们之所以在 CD-CAT 选题过程中，纳入项目区分度是出于两个原因：①在 Cheng（2009）提出 PWKL 时，同时也提出了 HKL，但 Cheng 自己的研究以及后续研究都指出 HKL 并没有更进一步的优势，这引起了我们长时间思考：有可能是加权方法的问题。②a 分层选题法另辟蹊径，将区分度摆到一个重要位置，其实从 IRT 中计算信

息量的公式里也可以看出区分度 a 的重要性。但在 CDA 中，并没有类似 a 的指标，使得先分层再选题的做法无法实现，但它给了我们灵感：在 CDA 中，项目区分度是否也具有同样重要的作用呢？HKL 是对个体层面的双加权指标，因为距离加权部分仍然利用的是个体层面的信息，那么在保留后验概率的基础上，将另一个个体层面的权重更换为项目层面的权重，是否会产生神奇的效果呢，于是我们就对该问题进行了研究。当然，我们也进行了只采用项目区分度进行选题的研究，结果并不理想，这表明单靠项目区分度选题的做法也是行不通的。于是，便有了将项目区分度作为权重的想法。

意见 2：因本研究是将项目区分度指标做为权重，请在文中就 CTT，IRT，CDA 这三种不同的测量理论下有关项目区分度的概念、联系与不同进行详细说明。

回应：感谢审稿人的意见。我们认真讨论该意见之后认为，本文的主要贡献是提出 CD-CAT 新的选题策略，而不是对项目区分度本身的研究，项目区分度是作为新指标的权重存在，因此，没有必要在文中专门对 CTT，IRT，CDA 这三种不同的测量理论下有关项目区分度的概念、联系与不同进行详细说明，这样会冲淡研究的主题与意义。因此，我们只把区分度的共性和主要作用在文中进行了阐述（请参见引言第二段的蓝色字体部分），将其他内容在此处进行回答，感兴趣的读者可以参见这部分。

①概念：

- CTT 中，项目区分度指题目对被试特质差异的区分能力。若所有考生在某个题目均答对或均答错，则此题目不能区分不同特质的被试，即此题目无区分能力。一般来讲，题目是否有区分能力，主要是看不同能力的被试在同一题目上得分是否不同。理论上，如果题目具有高区分力，那么高能力的被试在题目上应得高分，低能力被试在题目应得低分，也即被试的能力与在题目上的得分应存高相关。

- IRT 中，项目区分度 a 被定义为 $a=\sqrt{2\pi}\,tgA$，A 为项目特征曲线拐点的切线夹角，tgA 为其斜率，可见 a 是曲线拐点的斜率，决定曲线的陡峭程度。注意，在曲线拐点（难度）附近，曲线越陡峭说明不同 θ 被试正确作答概率的差异越大，因此将 a 定义为区分度参数。

  以上 CTT 与 IRT 区分度的概念均来自戴海琦主编的《心理测量学》（高等教育出版社）一书。

- CDA 中，项目区分度被定义为项目能够区别出不同知识状态被试的能力（discriminatory power of an item was defined as the ability of an item to differentiate between respondents with different attribute profiles）。该定义来自 Rupp 等（2010）一书 P304 页的内容。

②联系与区别

通过上述定义可以看出，虽然不同理论下，项目区分度的计算方法，敏感性等会有不同，但三者都是反映题目质量的指标，均是用来衡量题目能否有效区分出高能力被试和低能力被试（或不同知识状态）的关键指标。并且，1968 年洛德和诺维克在能力正态分布的假设下推导出了 IRT 中区分度 a 和 CTT 中区分度 r 的关系：$a=r/\sqrt{1-r^2}$，其中 r 是题目得分与

测验得分的双列相关系数。

意见 3：第“6 研究结论与讨论”部分，阐述不够清晰，请作者分“讨论”和“研究结论”两部分，分别进行恰当的、清晰阐述。
回应：感谢审稿人的意见，我们已经将第六部分分为“研究结论”和“讨论”两部分进行了阐述。

以下是论文中批注的修改意见：
意见 4：什么个体层面的信息，当个体进行项目作答时，实际也就利用了项目信息。就予以解释。
回应：我们已将“个体层面”的解释放在了脚注部分，即“个体层面是指 PWKL 指标在 KL 指标基础上融入的被试 KS 后验概率信息。”本文所说的“忽视项目本身能够提供的信息”是指选题指标中忽视了该部分信息，与被试在进行作答时利用了项目信息是不矛盾的。

意见 5：认知诊断中的项目区分度与 CTT，IRT 中的项目区分度有何区别？为何作者在此要区别指出是认知诊断中的项目区分度？
回应：因为这是在认知诊断理论下做的研究，所以要结合的是认知诊断中的项目区分度。关于 CTT，IRT 与认知诊断的项目区分度的概念，联系和区别请见对您的第二条意见的回复。

意见 6：模拟研究是在拟定的条件下获得的研究结论，针对实际运用，是否也是在相应的条件下，请作者清晰进行阐述。
回应：感谢审稿人的意见。我们已对该句话进行了修改：“因此，若实际测验情景与本研究的模拟情景相似，推荐 GIDPWKL 方法。”

意见 7：是否有参考文献，如有请备注
回应：该说法是我们基于自己的研究给出的一种新说法，是按照对 PWKL 加权方式的理解提出来的。这是因为 PWKL 在 KL 的基础上，将被试的后验概率信息作为权重加入了进来，因此，我们将之称作单源指标，这是合理的。

意见 8：Chang 和 Ying 提出的 a 分层法，不仅是基于 a 参数的实际作用，更是控制项目曝光率，提高题库的使用率。
回应：您的理解是正确的。a 分层法的作用是用来进行曝光控制，但这和本文的研究并不矛盾。引言部分之所以介绍 a 分层法，是想借用该方法来突出项目区分度的重要作用。

意见9：请就CDA中项目区分度的概念再加以清晰的表述。
（1）CDA与CTT，IRT中的项目区分度的异同

（2）这两类有关CDA的项目区分度的界定与CTT，IRT中有关项目区分度概念的联系与区别

这两类有关 CDA 的项目区分度界定的异同。
回应：感谢审稿人的意见。但和您第（2）条意见类似，本研究并不是对项目区分度本身的研究，而是在选题方法中，将项目区分度作为权重对 PWKL 进行了改进。如果要在论文对 CTT，IRT，以及 CDA 的项目区分度做专门介绍，会冲淡研究的主题与意义。我们在此对您的问题进行回复。

（1） 三种测量理论中的区分度含义其实是一样的，均是用来作为是否能够有效区别被试能力水平的指标。不同点在于，CTT 和 IRT 的潜在特质是连续的（前者是总分，后者是能力值θ），题目是对连续特质的区分；但 CDA 中项目是对离散知识状态的区分，因此，在 CDA 中，费舍信息量就不再适用了。

（2） 在 CTT 中，二值计分的项目区分度 dj 的一种求法为：dj = pu - pl。其中，pu 和 pl 分别是高分组和低分组的通过率，其思想是尽可能拉大真正能够答对被试与真正未能答对被试之间的差距，差距越大，项目区分度越大。但在 CDA 中，测评的结果是以多维离散潜在变量呈现，不像 CTT 那样能够基于总分找到高分组和低分组。因此，Rupp 等(2012)仿照 CTT 的思想，定义了 CDA 中的项目区分度：$d_j = p_{\alpha_h} - p_{\alpha_l}$。

其中，p_{α_h} 表示掌握题目 j 考察的较多属性的正确作答概率，p_{α_l} 表示掌握题目 j 考察的较少属性的正确作答概率。其含义是：“该项目区别掌握‘较多’属性被试和掌握‘较少’属性被试的能力”，因此，若一道题目能够较好地区分掌握所有属性和未掌握所有属性的被试，则 d_j 的值就会很大。这是第一类型的区分度。

第二类区分度是基于 KL 信息量提出的，因为在 CDA 中，潜在特质都是离散的，因此无法像 IRT 那样可以求得费舍信息量，但 KL 信息量不受具体分布的限制，所以在 CDA 中得到了广泛的应用。在 CDA 中，KL 值越大，表明该题目越能将不同的知识状态区分开。

（3） 从上一问题可以看出，CDA 中两类项目区分度构建的理论背景和出发点是不一样的。一个是沿用 CTT 的思想，将项目区分度定义为项目区别掌握较多属性被试和掌握较少属性被试的能力。另一个是基于 KL 信息量的角度，将项目区分度定义为项目区别不同分布，或被试的不同知识状态的能力。

对这些区分度的相关介绍还可以参考本文 3.1 至 3.6 部分。

意见 10：阅读上文，这并不是作者选择 DINA 模型的理由。
回应：感谢审稿人的提问。我们在相应部分增加了对 DINA 模型的描述，即“DINA 模型是认知诊断研究中最常使用的模型，由于 DINA 模型参数较少、简单易懂、方便解释，因此成为了许多研究者修正和拓展的基础模型。”

意见 11：是否有参考文献，如有请备注。
回应：该说法是我们基于自己的研究给出的一种新说法，因为本文的创新点之一就是在原有指标基础上增加新的信息源。之所以将新方法称作多源指标(multiple-source index)，是因为：PWKL 是在 KL 信息量基础上融入了被试的后验概率信息，因此 PWKL 被定义为单源指标(single-source index)，而新的方法是在 PWKL 基础上又加入了项目区分度信息，这点和 HKL 是不同的，所以我们将之称作多源指标，这是合理的。

意见 12：为什么会出现这样的结果，请给出适当的解释。
回应：说明这三种方法（CIDPWKL,KLEDPWKL,MIM）的表现相差无几，并没有表现出更大的优势。

意见 13：也就是说，随着测验长度的增加，GIDPWKL 和 AIDPWKL 这两种选题策略的精度增加的幅度越来越小，为什么会这样？

回应：因为随着测验长度的增加，被试提供的信息会越来越多，不同方法之间的差异是会呈现逐渐减小的趋势，这不仅是新方法存在的现象，是所有方法都有现象。我们用一篇论文中的具体实验结果来进行展示（如下表 3）。而且这也正是新方法优势的充分体现：测验长度越短，判准率的提升幅度越大。

表 3　CD-CAT 中不同测验长度下各选题策略的属性和模式判准率

测验长度	选题策略	属性 1	属性 2	属性 3	属性 4	属性 5	属性 6	整个模式
12	RAND	0.7910	0.8180	0.7990	0.7950	0.8050	0.8010	0.3170
	KL	0.9480	0.7400	0.9920	0.5490	0.6770	0.8430	0.2080
	SHE	0.9770	0.9660	0.9730	0.9720	0.9760	0.9760	0.8700
	PWKL	0.9720	0.9730	0.9730	0.9650	0.9680	0.9690	0.8600
	HKL	0.9780	0.9650	0.9760	0.9540	0.9660	0.9630	0.8540
16	RAND	0.8360	0.8470	0.8460	0.8450	0.8330	0.8550	0.4340
	KL	0.9600	0.8760	0.9960	0.6070	0.8100	0.9070	0.3640
	SHE	0.9920	0.9870	0.9900	0.9840	0.9940	0.9930	0.9430
	PWKL	0.9880	0.9860	0.9830	0.9890	0.9800	0.9830	0.9310
	HKL	0.9920	0.9860	0.9860	0.9850	0.9810	0.9920	0.9350
20	RAND	0.8700	0.8630	0.8810	0.8870	0.8690	0.8690	0.5070
	KL	0.9710	0.8970	**0.9870**	0.6630	0.8760	0.9100	0.4610
	SHE	0.9980	0.9990	0.9970	0.9970	0.9960	0.9960	0.9850
	PWKL	0.9940	0.9910	0.9960	0.9950	0.9960	0.9930	0.9680
	HKL	0.9950	0.9950	0.9960	0.9940	0.9940	0.9970	0.9750
24	RAND	0.9020	0.8900	0.8940	0.8910	0.9000	0.8970	0.5860
	KL	0.9860	0.9060	**0.9950**	0.7420	**0.8660**	0.9360	0.5490
	SHE	0.9990	1.0000	0.9990	0.9980	0.9960	0.9980	0.9930
	PWKL	0.9990	0.9980	0.9990	0.9990	0.9980	0.9960	0.9900
	HKL	0.9970	0.9990	0.9980	0.9990	0.9980	0.9970	0.9880

摘自：陈平. (2011). 认知诊断计算机化自适应测验的项目增补——以 DINA 模型为例. 北京师范大学博士学位论文.

意见 14：为什么 Q 矩阵结构的复杂性会影响选题策略的精确度。

回应：因为复杂 Q 矩阵结构中每道题目测量的属性比简单结构更多，在相同题目数量的情况下，复杂的 Q 矩阵更有可能考察到每个属性多次，这种情况类似于多级评分题目比 0-1 评分题目提供的信息更多一样。

意见 15：GIDPWKL 和 AIDPWKL 这两种方法的判准率较好，但是它们的项目曝光情况，相比较而言并不是很理想。

回应：审稿人的理解是正确的。这是因为在 CAT 中，估计精度和题库使用情况之间存在着权衡（trade-off）问题。若研究目的在于提升估计精度，那题库使用情况势必会受到影响。若更加注重题库的使用，估计精度又会受到影响，这在所有的 CAT 研究中都是普遍存在的现象。在这两者之间如何抉择，需要根据研究目的和实际使用情景来定。本文的研究目的及重点是在较短测验时，如何能够快速地提升对被试 KS 的估计精度，题库的使用会受到影响是可以预期的。根据研究结果可以看出，在提高判准率的同时，测验重叠率和卡方值比起 PWKL 的结果上升得并不多。另外，本文提出的方法是用于实际课堂教学中让教师快速地了解学生当前的知识掌握情况，因此，比起估计精度，题库的使用情况并不是本研究关注的

重点，而且本研究在进行实验时考虑了项目曝光控制技术，得到的题库使用结果是在预期控制范围内的。

意见 16：这部分没有对研究问题，研究设计和研究结果等展开深入的相关讨论；其次研究结论部分没有进行简洁清晰的阐述。请分讨论和研究结论两个部分分别进行恰当的、清晰的阐述。

回应：该问题同您提出的第（3）个问题一样。我们已经按照您的意见，将第六部分分为“研究结论”和“讨论”两部分进行了阐述。

第三轮

审稿人 1 意见：

经过两轮的修改，作者较好的吸收了几位审稿人的意见，特别是增加了 KLEDPWKL 和 MLM 方法的仿真，使文章内容和表述都有得到提高。

意见 1：但是，本文增加的讨论部分放在了第 7 节、而研究结论放在第 6 节，与常规的心理学论文写法不同。另外，第 6 节的第一段也有些讨论性的内容。是否可以考虑把这两部分重新整理和梳理一下。

回应：感谢审稿人对本文的肯定，以及提出的意见。

关于“结论”和“讨论”部分的写作格式，我们是遵循了第二轮第三位审稿人的第三条意见，将原本的“研究结论与讨论”拆成了两部分分别进行了描述，我们认为这样的描述也不失一般性。同时，我们阅读了本领域其他作者在心理学报发表的文章，有的作者会将结论与结果放在一起，最后再写讨论；有的作者将结论和讨论放在一起写作。我们认为您的意思是“结论”部分不应该出现讨论性质的内容，因此，我们删除了该部分讨论性的内容。

审稿人 2 意见：

文中将 CTT，CD 提出的项目区分度指标做为 PWKL 的权重，构成新的选题指标，并基于这些指标的表现进行了模拟比较研究。作者对前两次审稿意见都做了较好的回复与修改，通读后，有几处疑惑：

意见 1：引言部分作者提及项目曝光率的问题，令人费解作者想要表达什么。是在下文当中也将讨论项目曝光率问题吗？但是作者更多的是从判准率的角度来讨论各策略的优劣。

回应：引言部分提及a分层选题法只是将其作为例子，目的是为了突出项目区分度在测验中的重要作用。第三段提及曝光问题，是因为CAT的安全性已成为研究者关注的焦点之一，这不仅是理论上的需求，更是现实测验中必须要考虑的实际问题。许多研究者也专门对CD-CAT的项目曝光控制进行了研究：

1. 陈平. (2011). *认知诊断计算机化自适应测验的项目增补 ——以DINA 模型为例*(博士学位论文). 北京师范大学.
2. 郭磊, 郑蝉金, 边玉芳. (2015). 变长CD-CAT中的曝光控制与终止规则. *心理学报, 47*(1), 129-140.
3. Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*(7), 563–582.
4. 毛秀珍, 辛涛. (2013). 认知诊断CAT中项目曝光控制方法的比较, *心理学报, 45*(6), 694-703.

5. Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of EducationalMeasurement, 48*(3), 255–273.
6. 汪文义, 丁树良, 宋丽红. (2014). 兼顾测验效率和题库使用率的CD－CAT选题策略. *心理科学, 37*(1), 212-216.

因此，本文在进行模拟研究时，对项目进行了曝光控制，更加贴近真实测验情景。此处对曝光率问题进行介绍，是为了告诉读者本研究中要对项目曝光率进行控制，起到铺垫作用。

在 CD-CAT 的研究中，判准率是各种方法进行比较的最重要的指标。除此之外，本文也将题库使用情况作为了另一个比较指标。因此，文中所有的选题方法均是在判准率和题库使用情况这两个指标上进行的比较。

意见 2：从 4.4 的结果来看，虽然精度与题库使用情况是个权衡问题，文中用项目区分度指标来作为权重，但是最后针对题目的均匀使用、题目的曝光率问题并没有从中获得很好的解决，这是否也是单纯地从加权的角度来解决问题最后导致的结果。

回应：感谢审稿人提出的问题。使用项目区分度作为权重对 PWKL 指标进行改进，其目的是为了快速地提升测验初始阶段对被试 KS 的判准率，这是要以牺牲题库使用均匀性为代价的，并且在所有的相似研究中，这种情况都不可避免。因此，为了防止各种选题方法过度地使用质量较好的题目，我们在研究中才考虑加入曝光控制技术。从对第一轮第一位审稿人的第四条回答中可以看出，题目的曝光率全都低于允许的最大曝光率 0.2 以下。

意见 3：在“6 研究结论部分：另一方面，根据 Wang(2013)的研究结果表明，MIM 在大部分实验条件下的表现要优于 PWKL，特别是在测验长度较短时，但作者并未考虑在曝光控制条件下 MIM 的表现，也没有新方法与 MIM 之间的比较研究……”。作者提到没有考虑新方法与 MIM 之间的比较，但是下文又提及，系统比较了六种方法，这六种方法当中有 MIM 方法。不明白作者要说明什么。

回应：这是我们写作上的问题，抱歉给您带来了困扰。我们要表达的意思是 Wang 本人对 MIM 和 PWKL 进行了比较，发现 MIM 要优于 PWKL，但她并没有提出项目区分度的选题方法，因此，为了突显本文提出新方法的优势，必然要和 MIM 进行比较。另外，Wang 在她的研究中并未考虑曝光控制问题，这可能会与实际测验情景不符，因此，在本文中，我们还同时考虑了曝光控制。我们将文中的原句修改为：“另一方面，根据 Wang(2013)的研究结果表明：MIM 在大部分实验条件下的表现要优于 PWKL，特别是在测验长度较短时。但 Wang 本人并未考虑在曝光控制条件下 MIM 的表现，目前也没有新方法与 MIM 之间的比较研究。”

以下是论文中批注的修改意见：

意见 4：将区分度指标作为权重，或是一个单纯的数学公式上的考量因子，我个人并不认为这是一个好的、新的选题策略。在实际测量过程当中，除了数学上的考量之外，还需要针对实际情况对测验进行考虑，比如题库的结构、设计、比如实际作答的情况等等。完全脱离测量实际的、纯粹从数学上的去思考问题只是解决问题的一个较片面的思路。

回应：感谢审稿人提出的问题。“加权”在统计中是一种解决问题的重要思想。加权是为了：突出某些因素的作用，或平衡多个因素之间的关系。在心理测量学中，项目区分度的重要程度不言而喻，编制测验时，通常都想获得较高区分度的题目，区分度越高，该题目能够区分不同能力水平被试的效能就越大。因此，选题时，在原有方法的思想基础上，纳入对项目区分度的考量不仅突出了“区分度”的作用（这是当前CD-CAT选题方法中均未考虑的），还平

衡了原来方法只从个体层面信息进行加权的缺陷。因此，我们认为本文所提出的项目加权的选题方法并不是纯粹从数学上进行的考量，这些方法更是结合了心理测量学中区分度指标的优势所提出来的。相关的回答您还可以参见我们对第二轮第三位审稿人的第一条回答。同时，我们还可以从很多其他研究中看到加权的重要作用和重大意义。例如，Cheng等(2009)为了改进CD-CAT的选题效率，在KL的基础上进行了后验加权，从而提出了PWKL方法[Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. Psychometrika, 74(4), 619-632.]。Cheng等(2009)为了同时控制曝光、内容平衡、答案平衡等因素，提出了最大优先指标的加权方法[Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. British Journal of Mathematical and Statistical Psychology, 62, 369–383.]。Cheng等(2009)为了平衡测量精度和曝光，在a分层基础上进行了加权[Cheng,Y., Chang, H. H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints balancing measurement efficiency and exposure control. Educational and Psychological Measurement, 69(1), 35-49.]。Wang等(2012)为了同时兼顾能力值和KS的估计精度，提出了相应的加权选题方法[Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. Behav Res Methods, 44(1), 95-109.]等。

您提到的“还需考虑题库的结构、设计、实际作答的情况等等的相关问题”，我们在模拟研究中已经进行了考虑，包括 Q 矩阵结构，题目质量和测验长度等。当然，一篇论文不可能将所有情况都研究的很透彻，我们在“讨论与展望”部分对未来值得进一步研究的问题进行了展望。

我们认为任何科学研究都是具有探索性本质的。新的研究只要在理论上有据可依，逻辑清晰，并且通过模拟研究证明新方法比 PWKL 方法更具优越性，那么该方法就是好的、新的方法。

意见 5：在变长测验情景下，四种新方法和 MIM 的平均测验长度均要低于 PWKL 方法，是否存在显著性差异。

回应：我们查阅了已发表的认知诊断方面的论文，均未发现有做显著性检验的。我们是这样认为的：在进行方法比较时，并非像实验研究那样，一定要在实验处理之间达到显著性才能说处理有效。一个方法是否优于另一个方法，主要体现在判准率（定长）或平均测验长度（变长）上。在控制其他条件相同时，若使用某个方法能够得到更高的判准率或更短的平均测验长度，这种方法就是更佳的方法。

意见 6：是否已经用于实践。作者在此举例说明，实际当中似乎难以实现。首先 CD-CAT 要在机房进行，可我们国内大部分的课堂都不是在机房；其次，根据 CD-CAT 的诊断结果，每个个体的情况不一样，但是课堂教学是要面对一个集体，老师如何做到有针对性的教学；第三准确估计个体的知识状态，除了跟选题策略有关，还跟题目设计、认知诊断模型和方法等有关。

回应：感谢审稿人提出的问题。

①CD-CAT已经用于了实践，例如，北京师范大学的刘红云教授发表的论文《Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. Journal of Classification, 30(2), 152-172.》中，已对北京市8所学校的5年级和6年级一共584名学生进行了CD-CAT的施测；美国伊利诺伊大学香槟分校的张华华教授团队成功地研发了物理学科的

CD-CAT测验，在美国已经用于实际教学中，并且该团队正在将物理学科CD-CAT的成功经验借鉴到我国的HSK考试中；北京师范大学的陈平老师所在的项目组正在创建小学中、高年级数学“行程问题解决”领域的CD-CAT题库并正在筹备相关测验系统的研发。可以看出，CD-CAT的应用不仅能够更好地指导教师教学，更能促进学生的自我提升。

②当我们知道学生的知识漏洞后，可以将具有共同缺陷的学生组织在一起进行针对性的教学补救，提高教学效率。

③感谢您的意见，您说的很对，KS 的估计精度确实跟题目设计、认知诊断模型和方法有关，但本文聚焦的核心问题不是这些，我们在今后的研究中还会进一步探讨这些问题。

意见 7：PWKL 方法没有考虑区分度指标，是因为忽略，还是因为其它原因。不能因为 PWKL 没有考虑区分度指标，所以我们要考虑区分度指标。

回应：感谢审稿人提出的问题。PWKL 是由 Cheng（2009）提出的，我们猜测可能是她没有往区分度的角度进行考虑，所以没能提出项目区分度加权的选题方法。我们并非是因为 PWKL 没有考虑区分度而要去考虑区分度，相似的问题回答请参见我们对第二轮第一位审稿人第一个问题的回答，第二轮第三位审稿人第一个问题的回答，以及对您第一条批注的回答

意见 8：这种方法是本文作者提出，还是汪文义提出？

回应：KLEDPWKL 方法是本文提出的，之所以借用了 KLED 这个名字，是因为 KLEDPWKL 中的“权重”是来自汪文义等人的研究。

其余的问题我们直接在论文中相应部分进行了修改，用高亮字体标志了出来，因此就不在此一一赘述。

最后，感谢两位审稿人认真细致的审稿工作，谢谢。