《心理学报》审稿意见与作者回应

题目：变长 CD-CAT 中的曝光控制与终止规则

作者：郭磊 郑蝉金 边玉芳

第一轮

审稿人 1 意见：变长 CD-CAT 中曝光控制与终止规则（特别是终止规则）是一个研究结果很少的课题，因此开展这方面的研究很有意义。但是文章存在如下一些问题：

意见 1：引言中第二段的结论“目前，CD-CAT 的研究主要集中在选题策略、项目曝光控制; 毛秀珍, 辛涛, 2013)和属性在线标定方面。但不论以上哪方面的研究，均是以定长（fixed-length）CAT 的形式作为其终止规则，即固定每次测验的长度，当被试完成测验后，根据被试的作答反应估计知识状态”。这样关于“目前的研究现状”的叙述却和下文中所说的 Hsu,Wang & Chen(2013)的研究不完全符合，因为 Hsu 等人的研究就是变长 CD-CAT。而且关于属性在线标定的文献不多，但是综述时却遗漏了汪文义等人（2011）在心理学报上的研究。

回应：感谢审稿人对文章写作逻辑提出的意见。根据您的意见，我们已将正文第二段内容进行了逻辑上的修改，将 Hsu,Wang 和 Chen(2013)的研究作为目前 CD-CAT 研究的一部分进行了叙述，并新增加了两篇参考文献，分别是 Wang（2013）和汪文义, 丁树良, 游晓锋（2011）的文献。

意见 2：在文章的“结论与讨论”第二部分中说“Q 矩阵还可以包括线型、收敛型、发散型及它们组合起来的更为复杂的结构，而且，实际中的 Q 矩阵确实是错综复杂的”，这里似乎应该说“属性层级还可以包含括线型、收敛型、发散型及它们组合起来的更为复杂的结构”，这样表述可能更加准确。

回应：感谢审稿人的宝贵意见，我们已将 *Q* 矩阵的说法改为属性层级。

意见 3：文章 3.3.3 MRT 法一节说“该方法首先。。。”，在文章中有“首先”，但是再找不到“然后。。。”或者“其次。。。”的内容，文章是否有遗漏？

回应：感谢审稿人的宝贵意见，这是本人行文时的纰漏，现已将“首先”一词删除。

意见 4：没有说明为什么要将“RP 和 RT”修改为“MRO 和 MRT”；而且文章中公式（5）的修改和 RP 的完全不同，RP 是 Rj 和 PWKL β 的加权，加权系数（1-x/L）和 x/L,但是作者修改时，中括号里面加权系数发生变化，这可以理解；问题是修改以后大括号外面为什么还有一个“曝光控制因子 f_h ”，作者没有给出相应的解释；

回应：感谢审稿人的宝贵意见。我们在 3.3 部分加入了将“RP 和 RT”修改为“MRP 和 MRT”的原因阐述。文中公式（5）的修改理由在 3.3.2 部分进行了说明，公式中的加权系数 $(P_{1st}-P_{current})/P_{1st}=1-P_{current}/P_{1st}$，形式上和（1-x/L）是一样的。“曝光控制因子 f_h ”是加在大括号里面的。原因正如 Wang, Chang 和 Huebner（2011）文章中所述一样，如果只在 MRP 法中加入随机成分，不足以保证有效地控制所有过度曝光的项目，为了保证所有的项目均能得到控制，因此需要加入一个限制因子，即文中所说的“曝光控制因子 f_h ”，这也正是该方法中“Restrictive”一词的含义。本文沿用了 Wang, Chang 和 Huebner（2011）文中的做法，并且根据实验结果，取得了非常理想的结果。同时，我们在文中 3.3.2 部分加上了引入“曝光控制因子 f_h ”的理由。

意见 5：作者考察按照精度终止的被试模式判准率 PCCR(p)和按照最大长度终止的被试模式判准率 PCCR(max)。一方面，变长 CD-CAT 可以用比较短的测验得到比较准确的知识状态的估计，所以一般按照测量精度终止的测验的长度应该比按照最大长度终止的测验的长度要短；另一方面，一般来讲，测验长度越长，测验准确率越高；但是文章表 2,3,4 的不少数据表现出在其他试验条件相同的情况下，PCCR（p）甚至高于 PCCR（max），不知道作者注意到这个试验数据吗？如何解释？

回应：非常感谢审稿人给出此意见，让我们发现了原文写作不清的问题，这些问题已经在修改稿中进行了相应的修改和补充。我们认为测验长度与判准率之间的关系需要分情况讨论：在定长 CAT 情境下，测验长度越长，测验准确率会越高，这也是大部分定长 CAT 研究的一个结论。这里要求所有被试全部都做一样长度的题目，判准率是基于全体被试报告的，因此，做的题目数量越多，挖掘的信息也就越多，测验准确率就越高。但在变长 CAT 情景下，通常是根据不同的终止规则来设置期望要达到的标准，如果终止标准设置的合理有效，是不会出现按照最大测验长度终止的情况。但我们知道，在模拟与实际测验中，总会有部分被试需要作答很多项目才能达到预设标准（原因可能有多种：和选出来的题目，题库质量与题目质

量、终止标准等都有关系），甚至需要做完题库中所有的项目，但我们不能让被试无休止地做下去，因此有必要设定一个合理的最多答题数（例如 30 题），提前终止 CAT，这也是很多变长 CAT 研究中采取的常用方法，而且这些研究得到的结果均是按照精度停止的判准率要高于按照测验长度停止的判准率。

我们在认真思考表 2,3,4 出现的问题之后，认为有两个观点需要厘清：一个是 PCCR（p）高于 PCCR（max）的情况，另一个是 PCCR（max）高于 PCCR（p）的情况。第一种情况很好理解，是因为在理想情况下，一旦被试 KS 的判准精度达到预设水平后，CAT 就结束。利用最大长度终止测验的被试比利用预设标准终止测验的被试的判准率要低，这是因为利用最大长度终止规则停止的那些被试始终未能达到预设的精度标准，因此 PCCR（p）是要高于 PCCR（max）的，这与 Hsu, Wang 和 Chen (2013)的研究结果是一致的；出现第二种情况（即 PCCR(max)有时候会高于 PCCR(p)）则是由模拟的随机性所导致的。为了解决这个问题，我们重新模拟了所有实验条件，并且在每个实验条件下都重复 30 次，以此来减少随机误差。表 1 至表 4 的结果全都是 30 次模拟的平均结果。下面以其中几个比较重要的结果为代表说明该问题：

1） 表1-1是HSU法（P_1=0.8）在MRT曝光控制下30次的实验结果（对应正文表3第1行的实验条件）。从30次实验结果可以看出，PCCR(p)的估计是很稳定的，基本上围绕在平均值（0.69）周围浮动；PCCR(max)的结果很不稳定，这些被试（%max的平均值为0.2%，即2000*0.2%=4人）未能按照精度停止，导致了他们知识状态（KS）估计的不可控性，即，这时他们做完30题后精度仍然没有达到要求。由于模拟研究存在一定程度的随机性，当这4名被试做到第30题时，KS恰好都估计正确的可能性是存在的，KS恰好都估计错误的可能性也是存在的，因此，PCCR(max)的结果会浮动的比较大，这是合理的。并且，根据Choi, Grady和Dodd（2010）在A New Stopping Rule for Computerized Adaptive Testing [EPM] 文中的描述：“… in some cases, meaningful gains in measurement precision may be both desirable and possible with the administration of only one or two additional items.”（PP. 39），我们可知，在变长CD-CAT中的道理是一样的。当被试做到30题时（即按照最大测验长度终止测验），相比变长测验的平均情况而言，大约多做了17题，根据Choi等人的观点，被试多做几道题目是会存在增加估计精度可能性的，因此从这个观点出发，也可解释为什么PCCR(max)有时会等于1，或者是超过PCCR(p)。但将30次实验结果求平均之后，PCCR(max)的均值为0.69，要小于PCCR(p)的均值0.86。原文中的0.9999只是一次实验的结果，具有偶然性，因此在返修稿中，我们将结果改为了30次实验的平均值。30次

实验的结果如表1-1所示，并且在正文5.2部分中也加入了对该部分的分析与讨论。

表 1-1 30 次实验结果汇总（HSU）

实验次数	1	2	3	4	5	6	7	8	9	10
PCCR(max)	0.33	NaN	NaN	0.83	0	0.50	0.73	1.00	1.00	1.00
PCCR(p)	0.87	0.87	0.87	0.86	0.86	0.85	0.87	0.87	0.87	0.86
实验次数	11	12	13	14	15	16	17	18	19	20
PCCR(max)	0.50	0.33	0.75	NaN	1.00	NaN	1.00	0.67	0.67	1.00
PCCR(p)	0.86	0.85	0.87	0.86	0.86	0.86	0.86	0.87	0.87	0.87
实验次数	21	22	23	24	25	26	27	28	29	30
PCCR(max)	0.67	1.00	1.00	1.00	0.33	1.00	NaN	0.50	0.57	0
PCCR(p)	0.86	0.87	0.86	0.87	0.86	0.86	0.87	0.86	0.86	0.87

PCCR(max)的均值为 0.69；PCCR(p)的均值为 0.86

2） 表 1-2 是 DAPP 法（e=0.05）在 simple 曝光控制下 30 次的实验结果（对应正文表 2 第 23 行的实验条件）。从 30 次实验结果可以看出，PCCR(p)的估计是很稳定的，基本上围绕在平均值（0.34）周围浮动；PCCR(max)的结果很不稳定，解释同上。30 次实验的结果如表 1-2 所示：

表 1-2 30 次实验结果汇总(DAPP)

实验次数	1	2	3	4	5	6	7	8	9	10
PCCR(max)	1.00	NaN	0.89	1.00	0.80	1.00	0.67	0.00	0.00	0.86
PCCR(p)	0.32	0.33	0.31	0.33	0.33	0.32	0.32	0.34	0.31	0.33
实验次数	11	12	13	14	15	16	17	18	19	20
PCCR(max)	0.50	1.00	1.00	NaN	1.00	1.00	1.00	0.83	0.67	1.00
PCCR(p)	0.33	0.33	0.31	0.33	0.32	0.33	0.33	0.33	0.33	0.34
实验次数	21	22	23	24	25	26	27	28	29	30
PCCR(max)	1.00	1.00	1.00	1.00	0.86	1.00	0.67	1.00	1.00	0.67
PCCR(p)	0.34	0.32	0.31	0.33	0.32	0.31	0.34	0.34	0.33	0.34

PCCR(max)的均值为 0.83；PCCR(p)的均值为 0.34

3） 表 1-3 是 KL 法（e=0.05）在 MRT 曝光控制下 30 次的实验结果（对应正文表 3 第 27 行的实验条件）。从 30 次实验结果可以看出，PCCR(p)的估计是很稳定的，基本上围绕在平均值（0.37）周围浮动；PCCR(max)的结果很不稳定，解释同上。30 次实验的结果如表 1-3 所示：

表 1-3 30 次实验结果汇总(KL)

实验次数	1	2	3	4	5	6	7	8	9	10
PCCR(max)	NaN	NaN	NaN	1.00	1.00	NaN	0	1.00	NaN	NaN
PCCR(p)	0.36	0.37	0.35	0.37	0.38	0.36	0.37	0.35	0.34	0.34
实验次数	11	12	13	14	15	16	17	18	19	20
PCCR(max)	NaN	NaN	NaN	1.00	NaN	0	1.00	NaN	NaN	NaN
PCCR(p)	0.35	0.36	0.36	0.36	0.34	0.34	0.35	0.35	0.37	0.36
实验次数	21	22	23	24	25	26	27	28	29	30
PCCR(max)	1.00	0	0	1.00	0	NaN	NaN	NaN	NaN	NaN
PCCR(p)	0.35	0.36	0.36	0.36	0.36	0.37	0.35	0.35	0.36	0.35

PCCR(max)的均值为 0.58；PCCR(p)的均值为 0.37.

4） 表 1-4 是混合法 HM（P1=0.8 & e=0.05）在 simple 曝光控制下 30 次的实验结果（对应正文表 2 第 5 行的实验条件）。从 30 次实验结果可以看出，PCCR(p)的估计是很稳定的，基本上围绕在平均值（0.94）周围浮动；PCCR(max)的结果很不稳定，解释同上。30 次实验的结果如表 1-4 所示：

表 1-4 30 次实验结果汇总(HM)

实验次数	1	2	3	4	5	6	7	8	9	10
PCCR(max)	0.75	0.87	0.72	0.74	0.67	0.81	0.87	0.69	0.79	0.62
PCCR(p)	0.95	0.95	0.92	0.94	0.95	0.95	0.94	0.95	0.95	0.95
实验次数	11	12	13	14	15	16	17	18	19	20
PCCR(max)	0.88	0.90	0.96	0.91	0.82	0.69	0.76	0.80	0.72	0.64
PCCR(p)	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95
实验次数	21	22	23	24	25	26	27	28	29	30
PCCR(max)	0.75	0.56	0.72	0.86	0.62	0.88	0.73	0.77	0.81	0.65

PCCR(p)	0.94	0.94	0.95	0.94	0.95	0.95	0.94	0.94	0.94	0.95

PCCR(max)的均值为 0.77；PCCR(p)的均值为 0.94

尽管表 1-4 全部改成 30 次实验结果的平均值了，但我们发现在加入曝光控制之后，表 2 至表 4 中仍然有一些结果是 PCCR(max)要高于 PCCR(p)的（如粗体部分所示），这主要集中在 DAPP 法和 KL 法，其余方法并未出现这种矛盾结果，但这是符合我们预期的一种结果，由于我们的疏忽，没能在文章中对此问题进行讨论和说明，我们在修改稿中的结果与讨论部分加上了相关内容的阐述。以下是对出现这种情况的分析：出现这种情况的原因主要是由这两种方法的终止原理所导致的。DAPP 和 KL 法属于相对标准的终止规则，比起绝对标准的终止规则来说，它没有设置一个最低的临界标准，例如，最大后验概率至少要高于 0.8。也就是说，这两种方法完全有可能在 0.8 之前就符合前后之差低于预设水平 epsilon，甚至可能在最大后验概率很低时（例如，p1=0.4）就已经符合相对标准而停止测验了，这一点可以从平均做题量得到佐证。例如，表 2 中的 DAPP 法（e=0.05）的结果，所有被试的平均做题量只有 5.7 题，如果一个被试只做了大约 6 道题目，那么对其 KS 的估计应该是不够准确的，这表明曝光控制方法会影响到相对标准终止规则的表现。

意见 6：文章缺少一个总体的试验设计，比如多少因素，各个因素的水平数是多少，等等，没有描述，建议增加这一部分内容。

回应：感谢审稿人的宝贵意见。我们在 4.3 部分的结尾增加了总体的实验设计，包括因素的数量及各个因素的水平数，以及总体实验次数等内容。

审稿人 2 意见：

意见 1：建议删除摘要里的人名

回应：已删除。

意见 2：读者还未阅读该文献，如何知道 P1st 和 P2nd 是什么呢？建议将其替换为相对应的中文

回应：已替换为相应的中文名称。

意见 3：关于错别字，方法名字，参数含义，参考文献，斜体字等内容的问题

回应：我们已按照您的宝贵意见，在相应位置进行了修改或删除。

意见 4：关于正确作答概率和随机数比较大小模拟作答反应的问题。

回应：感谢审稿人提出的宝贵意见。原文中“大于”一词是本人在写作时由于疏忽大意，将“等于”一词漏掉了。我们程序中正是采用您说的“大于等于”方法模拟的。并且在此修改稿中，我们也将各个实验进行了 30 次的重复模拟，以此减少随机误差的干扰，使得最终结果更加可信。

第二轮

审稿人 1 意见：文章经过修改以后，比以前有进步。但是审稿人认为还存在以下 5 个问题

意见 1：逻辑上还是存在问题：在引言第二段最后面说“（本文将其称为 HSU 法，详见第 2 部分）其研究结果表明，随着知识状态后验分布中最大后验概率预设水平的升高或第二大后验概率预设水平的降低，被试的模式判准率均会上升，这是对变长 CD-CAT 研究的一大推动。”这里使用“或”字。

然而在 2.1 介绍 HSU 法时候说

Tatsuoka(2002)给出了变长 CD-CAT 的经验性准则，即被试属于某种知识状态的最大后验概率超过 0.8 时，测验终止。Hsu 等人（2013）基于 Tatsuoka 的思想，进一步提出了双重标准的变长 CD-CAT 终止规则，即当被试属于某个知识状态的最大后验概率 P_{1st} 不低于某个预设水平（例如，0.7），并且第二大后验概率 P_{2nd} 不高于某个预设水平（例如，0.1）时，测验终止。这里使用“并且”两个字。

请注意“或”字和“并且”在逻辑上完全不相同。

回应：感谢审稿人细心地阅读。“或”字的使用确实存在逻辑问题，我们已将引言第二段最后面的说法进行了修改，具体分为了两种情况，如修改稿中所述：“其研究结果表明，当固定知识状态后验分布的最大后验概率预设水平时，被试的模式判准率会随着第二大后验概率预设水平的降低而增大；当固定知识状态后验分布的第二大后验概率预设水平时，被试的模式判准率会随着最大后验概率预设水平的升高而增大。”

意见 2：（5）的左边的下标有问题，因为右边对 h 取得最大，已经和 h 无关。

回应：感谢审稿人的意见。我们已将公式（5）进行了修改。

意见 3：对于 2.4 节， 邻近后验概率之差法（difference of the adjacent posterior probability method, DAPP）

随着 CD-CAT 的进行，被试作答反应能提供的信息量越来越多，他属于某个“真实”的知识状态的后验概率会越来越大（Cheng，2009），（这是一个过程，何时达到最大）将此后验概率记为最大后验概率。（到底对什么取最大？）DAPP 法的测验终止规则为：当前后两次邻近的，并且是从属于同一种知识状态的最大后验概率之差的绝对值小于预设水平时，即 $\left|P_{t+1}(\hat{\alpha}_i)-P_t(\hat{\alpha}_i)\right|<\varepsilon$（$t$ 表示被试作答完 t 题，$P_t(\hat{\alpha}_i)$ 表示作答完 t 题后，KS 为 $\hat{\alpha}_i$ 对应的最大后验概率），测验终止。

请问如何体现“同一种知识状态的最大后验概率之差的绝对值小于预设水平时，即 $\left|P_{t+1}(\hat{\alpha}_i)-P_t(\hat{\alpha}_i)\right|<\varepsilon$（$t$ 表示被试作答完 t 题，$P_t(\hat{\alpha}_i)$ 表示作答完 t 题后，KS 为 $\hat{\alpha}_i$ 对应的最大后验概率），测验终止”是不是用 $\max\left|P_{t+1}(\hat{\alpha}_i)-P_t(\hat{\alpha}_i)\right|<\varepsilon$ 表示？如果是，如何定义上文的“最大后验概率”？

回应：感谢审稿人的意见。首先，被试真实知识状态的后验概率的变化确实是一个过程。不妨假设被试真实知识状态为（00101），当被试作答项目数量越多时，从他的作答反应中得到的信息就越多，那么与知识状态（00101）所对应的后验概率就会逐渐增大，我们认为这可能是您所说的“过程”的意思。其次，何时能达到最大值需要根据项目性能和作答反应来确定，一般而言，真实知识状态的后验概率会随着题目数量的增加而增大，但确定的数量关系目前还未有研究者研究过，同时，这也并非是本文关注的重点。本文是对终止规则的研究，因此，只要该后验概率满足其终止标准，就可以停止测验了。当停止测验时，我们可以查看平均用题量，平均用题量可以在一定程度上代表您所说的“能达到的最大值”。第三，关于“到底对什么取最大”，可能是我们写作没有交代清楚，导致了您的误解。DAPP 法中不涉及“取最大”的问题，“最大后验概率”中的“最大”一次是定语，是指被试知识状态后验分布中的最大的那个后验概率。DAPP 法是将前后两次最大的后验概率作差，然后取绝对值，再加以判断的一种终止方法。为了避免读者误解，我们根据您的意见，将 $\left|P_{t+1}(\hat{\alpha}_i)-P_t(\hat{\alpha}_i)\right|<\varepsilon$ 修改为 $\left|P_{1st}^{t+1}(\hat{\alpha}_i)-P_{1st}^{t}(\hat{\alpha}_i)\right|<\varepsilon$。

意见 4：对于 2.6 混合法（hybrid method, HM）

根据 Hsu 等人（2013）的研究结果可知，如果只控制 P_{1st} 不低于某个预设水平而未对 P_{2nd} 加以限制的话（即 Tatsuoka (2002) 的准则），被试知识状态估计的精确性并不理想。HM 法的测验终止规则为：当 P_{1st} 达到预设水平之后，再结合 DAPP 的做法，使得 $\left|P_{t+1}(\hat{\alpha}_i)-P_t(\hat{\alpha}_i)\right|<\varepsilon$（$t$ 表示被试作答完 t 题，$P_t(\hat{\alpha}_i)$ 表示作答完 t 题后，KS 为 $\hat{\alpha}_i$ 对应的最大后验概率）成立，测验终止。

注意 P 的两个下标 t+1 和 t，这么两个 P 都是最大后验概率吗？如何计算？

回应：这里的两个 P 都是最大后验概率，P_{t+1} 对应作答完 t+1 题后的最大后验概率，P_t 对应作答完 t 题后的最大后验概率。后验概率的计算公式为 $\pi_{i,t}(\alpha_c)=\dfrac{L(x_i^t\mid\alpha_c)\cdot\pi_{i,0}(\alpha_c)}{\sum_{l=1}^{2^K}L(x_i^t\mid\alpha_l)\cdot\pi_{i,0}(\alpha_l)}$（$c=1,2,\ldots 2^K$）。计算完所有知识状态的后验概率后，其中的最大值便是最大后验概率，记作 P_{1st}。

意见 5：文章中说“由于模拟研究存在随机性，加上按照最大测验长度终止测验缺少把握性，当被试做到第 30 题时，KS 恰好都估计正确的可能性是存在的，KS 恰好都估计错误的可能性也是存在的，因此，PCCR(max)值有较大的浮动是合理的”。

请问这里的“都”字是否意味着所有被试都估计正确（错误）？这种可能性（概率）到底有多大？是不是小概率事件？请作者再做试验，如果还出现这种情况，说明这不是小概率事件，应该好好查一查原因。

回应：感谢审稿人的意见。首先，%max 的比例跟很多因素有关，包括题库数量及质量，被试作答的随机性，允许的最大测验长度，所选用的认知诊断模型，终止精度等。理论上，当实验条件设置合理时，并不会出现%max 的情况，因为变长测验的优势就在于被试能够按照测验精度终止测验。其次，%max 的出现，确实是属于小概率事件，因为大部分实验条件下的%max 比例都是很低的（除了终止标准很严格时），表明绝大多数的被试是按照测验精度停止测验的。出现%max 并不会对本文得到的主要结果和结论产生冲击，这与 Hsu 等人的研究结果是一致的。第三，%max 比例较低，表明按照测验长度终止的被试人数是很少的，这部分被试构成了小样本，我们知道，对于小样本的估计本身就是不准确的，因此，30 次重复实验的结果存在差异是可以理解的。第四，我们在认真讨论和斟酌后，认为 30 次重复实

验的结果放在正文中不太合适，这给您及可能会给将来的读者造成误解，因此，我们将表 5 的结果从原文中进行了删除，只作为对第一位审稿人的回复。同时，我们增加了对相对标准终止规则出现 PCCR(max)高于 PCCR(p)的结果的探讨，请参见蓝色字体部分和图 1 至图 4。

最后，再次感谢您对本文的认真审阅和宝贵意见，谢谢您！

第三轮

意见 1：引言中说“Kingsbury 和 Houser (1993)的研究表明，不管是在多级评分 CAT 还是 0-1 评分 CAT 中，变长 CAT 在测验效率、能力估计的收敛和能力估计精度等方面均优于定长 CAT。相对于定长 CAT，变长 CAT 更能体现出自适应的特点和优势，因此，开展这方面的研究非常有意义”。结论与展望中又说“。。。而自适应测验的精髓应该是使得 CD-CAT 测验对每个被试的知识状态估计拥有相同的估计精度”。

审稿人认为引言所说的是针对基于项目反应理论（IRT）的 CAT 而言，因为 IRT 有 Fisher 信息量度量测量精度，可惜认知诊断理论中缺乏这样可以表示测量精度的信息量。这是变长 CD-CAT 的研究异常困难的根本原因。纵使作者的指标看上去和测量知识状态的精度有关，但是直觉不等于理论。至少作者的文章无法表明这一点。所以审稿人认为，研究变长 CD-CAT 的理由的阐述是不充分的。建议作者对此再研究。

回应：非常感谢审稿人对CD-CAT中的测量精度提出的问题。您说道：“IRT有Fisher信息量度量测量精度，可惜认知诊断理论中缺乏这样可以表示测量精度的信息量。”您说的很对，这也是目前CD中存在的一个困境。

首先，CD关注的是多维的离散变量，IRT关注的是连续的单维（或多维）变量，两者在理论基础上是存在本质差异的。IRT发展较早，对它的研究非常成熟，而CD的发展较晚，CD-CAT的出现就更晚了，因此，CD-CAT的理论在现阶段还没有达到IRT的成熟度。尽管目前还不清楚在CD中信息量与标准误之间的转换关系，但随着研究者的不懈努力，是有可能找到这种转换关系的，这种转换关系有可能是基于信息量角度的，更有可能是基于CD本身特点的另外一种指标体系。当这种关系被找到后，一定会大大推动CD-CAT整个领域的研究。再次感谢审稿人提出的这个问题，我们认为这不是变长终止规则本身的问题，而是CD-CAT整个系统的问题。

其次，本文的关注点正是CD-CAT领域在当前国际上的一个重点。C. Tatsuoka是CD研究的先驱人物，他于2002便经验性地提出了最大后验概率至少要高于0.8的变长终止准则（Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical*

Society: Series C (Applied Statistics), 51(3), 337-350.）。时隔11年后，Hsu等（2013）在Applied Psychological Measurement（心理统计与测量的权威杂志）上发表了双重标准的变长终止规则，这是基于Tatsuoka方法的一种改进（Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*(7), 563-582.）。另外，还有很多对CD-CAT研究的学者也在其文章的讨论部分提到了研究变长的重要意义，例如Cheng（2008）的博士学位论文，Wang（2011）在Journal of Educational Measurement（该杂志也是心理统计与测量的权威杂志）发表的文章的Discussion部分："However, variable-length testing is becoming more popular because of its "individualized" nature, which provides each examinee with roughly the same level of measurement precision. How to apply the current two methods to the variable-length CD-CATs is also a worthy topic for future research."因此，我们认为对变长终止规则的研究是非常有必要和有价值的。

第三，Hsu等（2013）文中说到："The rationale was that the more peaked the posterior distribution was, the more reliable the classification (Huebner,2010). In other words, posterior probability can be treated as a measure of precision." 我们非常同意该观点，也就是说，"精度"在IRT-CAT中是以能力估计的标准误来衡量的，但在CD-CAT中应该用后验概率分布来衡量。因此，本文基于最大后验概率（P_{1st}）提出了DAPP法和HM法，这两种方法的提出是有依据的。另外，每个知识状态KS是一个向量，无法直接获得向量的测量标准误。但是属性掌握的边际概率是可以求得标准误的。Rupp, Templin和Henson（2010）出版的*Diagnostic measurement: Theory, methods, and applications*（该书获得了AERA的大奖，也是目前国外教授认知诊断的指定使用教材）书中P242页给出了属性测量标准误的计算公式 $SE(\alpha_k)=\sqrt{P_k(1-P_k)}$。因此我们借用IRT-CAT的思想，提出了CD-CAT中的SEA法，该方法是和测量标准误直接相关的，可以直接反映测量精度，因此，该方法的提出也是有依据的。HA法的思想更加巧妙，它的思想来源于IRT-CAT的做法。在IRT-CAT中，有一种变长终止规则叫做"最小信息量终止规则"：即剩余题库中所有项目的项目信息量都低于某个预设水平时，测验结束（Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied psychological measurement, 19*(1), 5-22.）。而Tatsuoka最初提出HA法是作为选题策略使用的，但仔细分析HA法会发现，它也是跟最大后验概率有关系的一个指标。假设最大后验概率落入了某一个子集中，那么该子集的后验概率之和会增大，它的补集的后验概率之和会减小，两者的乘积项会减小，乘积项的最大值出现在最不确定时：

0.5*0.5=0.25。被试每多做一道题目，在理想情况下，乘积项会降低一些，当剩余题库中的所有题目无法再提供能让被试的后验概率乘积项降低的二分信息时，测验便终止，这正是HA法思想的巧妙所在，也为该方法找到了理论上的依据。因此，本文提出的方法是建立在现有的研究基础上，是对现有研究的推进，不是直觉的，是有理论支撑的。

第四，如前所述，虽然在CD中不存在信息量和标准误之间的转化关系，但是本文提出的方法作为变长终止规则使用是有较好表现的，读者可以从模拟结果中获得证实。

综上所述，CD-CAT 发展较晚，是目前测量领域中算是一个全新的研究领域。本文正是基于国际上的研究前沿进行研究的，就目前的研究结果来看，本文还是有新的贡献的：不仅提出了更丰富的变长终止规则，并且也提出了适用于变长测验情景下的曝光控制方法。假如今后在 CD 理论上有突破的话，我们会根据新的理论提出新的方法，进而推动该领域的研究。

意见 2：结论与展望中说“第二，本研究假设属性之间是无属性层级结构的关系”，这种说法有误。文章说“本研究题库及被试的知识状态采用陈平等(2011)的方法生成。陈平等(2011)在假设属性之间相互独立前提下”，独立结构本身就是一种属性层级关系，而不能说无属性层级结构的关系。

回应：感谢审稿人的意见，我们已经对其进行了修改，并用高亮表示。

第四轮

意见 1：正如作者所说“我们认为对变长终止规则的研究是非常有必要和有价值的”。审稿人十分赞成这一句话。正因为这一点，审稿人认为作者的研究有价值。

回应：非常感谢您对本研究的赞同！我们会继续做得更加深入。

意见 2：作者说“属性掌握的边际概率是可以求得标准误的”，因为边际属性是 0-1 随机变量，其方差等于成功概率（P）乘以失败概率（1-P），所以标准误立即可以求出。但是属性掌握模式（即知识状态）的方差矩阵是什么，这个矩阵和边际属性的方差的关系如何，文章没有交代。所以作者的研究离开目标还有一点距离。K K Tatsuoka 和 M M Tatsuoka 的儿子 C Tatsuoka （2002）提出的最大后验概率至少要高于 0.8 的变长终止准则的确有一定的道理，但是没有进行深入的理论阐述。因此审稿人认为文章作者在结论与展望中所说的“。。。而自适应测验的精髓应该是使得 CD-CAT 测验对每个被试的知识状态估计拥有相同的估计精度”还仅仅是一个良好的愿望，由于边际属性估计精度和知识状态估计精度的关系没有理清

楚，文章并没有实现这个愿望，因为纵使控制了边际属性的估计精度，也不能控制了知识状态的估计精度，也就不能够说知识状态估计拥有相同的估计精度。建议作者在结论和展望中讲清楚这一点。

回应：感谢您提出的宝贵观点和意见，我们将在讨论部分阐明我们对您提出的两个问题的思考。在这里我们给出更加详尽的说明。 首先，您指出了一个非常重要的问题，即知识状态方差（矩阵）的推导。实际上，认知诊断模型中知识状态方差的推导是一个非常困难的统计基础理论问题，如果有了知识状态的方差，我们就可以直接从知识状态的角度提出更加直接的终止规则，而不必从属性的边际分布间接的提出终止规则。C. Tatsuoka （2002）提出了目前最复杂、比较新的认知诊断理论之一，部分有序集理论（poset theory or partially ordered set theory），但是即使是这个模型也未能提供知识状态方差的推导。如果能够的话，我们相信他会直接提出基于方差的终止规则，而不是这个简单的 0.8 标准。本文的研究问题是 CD-CAT 的终止规则，是心理计量学中的一个研究问题，基本的定位是把先进的统计学原理及方法应用到心理测量中，因此推导知识状态的方差（矩阵）远远超出了本文的研究定位与目标。当然，如果能够首先实现统计基本理论问题的突破，然后基于新的统计理论研究结论提出新的终止规则，那将是比本文更加具有理论深度的研究，是我们未来的努力目标。

第二，我们将按照评审的要求在讨论部分指出标准误指标的优势与缺陷。我们提出的属性边际误差法（即 SEA 法）是利用边际分布对联合分布进行最大可能的逼近（approximation）。这种思路并不是我们的独创，在统计中广泛使用。在心理计量学中最经典的例子就是在结构方程(SEM)中用低维的边际分布来对高维的联合分布进行逼近（approximation），大大简化了建模与参数估计的难度。按照研究的数学推导，联合分布中的数学属性在边际分布中往往不能保持，但是这个数学推导的缺陷并没有严重损坏结构方程的科学性与实用性，阻碍其成为心理与教育以及其他社会科学最重要的统计工具之一。同样道理，我们的属性边际标准误指标也是借鉴了这样的思路。通过控制每一个属性的边际标准误（加上后验分布概率和为 1 的限制），我们间接地控制了整个知识状态后验分布。正如您指出的一样，这样的控制并没有完全实现“每个被试的知识状态估计拥有相同的估计精度”的理想状态，但是我们的方法是在现有认知诊断理论前提下，借鉴统计中利用边际分布估计联合分布的思想，尽可能去接近这个理想状态，并且实证研究表明我们的方法优于 Tatsuoka 提出的准则。我们将在讨论部分阐明这些优劣，让读者更加了解我们的思路。希望我们的回答能够让您满意。

意见 3：同意发表。

回应：再次感谢您对本文的肯定！

第五轮　编委复审意见

意见 1：请作者根据审稿专家如下的意见进行修改：

作者说“属性掌握的边际概率是可以求得标准误的”，因为边际属性是 0-1 随机变量，其方差等于成功概率（P）乘以失败概率（1-P），所以标准误立即可以求出。但是属性掌握模式（即知识状态）的方差矩阵是什么，这个矩阵和边际属性的方差的关系如何，文章没有交代。所以作者的研究离开目标还有一点距离。K K Tatsuoka 和 M M Tatsuoka 的儿子 C Tatsuoka（2002）提出的最大后验概率至少要高于 0.8 的变长终止准则的确有一定的道理，但是没有进行深入的理论阐述。因此审稿人认为文章作者在结论与展望中所说的“。。。而自适应测验的精髓应该是使得 CD-CAT 测验对每个被试的知识状态估计拥有相同的估计精度”还仅仅是一个良好的愿望，由于边际属性估计精度和知识状态估计精度的关系没有理清楚，文章并没有实现这个愿望，因为纵使控制了边际属性的估计精度，也不能控制了知识状态的估计精度，也就不能够说知识状态估计拥有相同的估计精度。建议作者在结论和展望中讲清楚这一点。

编委的答复：

第四次审稿人意见说“由于边际属性估计精度和知识状态估计精度的关系没有理清楚，文章并没有实现这个愿望，因为纵使控制了边际属性的估计精度，也不能控制了知识状态的估计精度，也就不能够说知识状态估计拥有相同的估计精度。建议作者在结论和展望中讲清楚这一点。”也就是说，审稿人希望作者在修改时，将“边际属性估计精度和知识状态估计精度的关系梳理清楚”。作者在修改时，确实对此进行了努力，在“2.2 褐色部分”，作者写道“理论上，应该根据知识状态的后验概率分布求取知识状态的方差，进而得到基于被试 KS 的标准误，这样就能实现对 KS 估计精度的直接操作。但可惜的是，我们无法直接求取 KS 的标准误。尽管整个模式的标准误无法求得，但单个属性的标准误是可以求取的。Rupp, Templin 和 Henson(2010; P242)书中给出了计算属性标准误的方法。”

但是，这一段话只是表示作者“无法直接求取 KS 的标准误”，也就是说，“整个模式的标准误无法求得”。作者在这里并没有回答审稿人提出的问题，而只是说自己无法得到知识状态的估计精度。

在“结论与展望”部分，作者写道“首先，SEA 法未能直接对知识状态的标准误进行操作，而是通过控制每一个属性的边际标准误来间接地实现对知识状态后验分布的控制。尽管这种利用低维的边际分布对高维的联合分布进行逼近（approximation）的方法未能保持联合分布中的数学属性，但这种做法在统计中有着广泛的使用，其中结构方程模型（*Structural Equation Modeling*, SEM）的参数估计就是最经典的例子，这种逼近方法并没有严重损坏 SEM 的科学性与实用性。当然，未来的研究需要进一步提出全新的认知诊断理论，推导出知识状态的方差，提出更加直接的 CD-CAT 变长终止规则。”

在这一段话中，作者只是以结构方差模型为例，指出“尽管这种利用低维的边际分布对高维的联合分布进行逼近的方法”是有缺陷的，但“并没有严重损坏 SEM 的科学性与实用性 ”。作者在这里只是运用举例的方法进行了解释，也没有直接回答审稿人的问题，即“边际属性估计精度和知识状态估计精度的关系”究竟是什么？如果把这个问题说清楚了，那么作者提出的这些新方法也就有了更加坚固的基础。因此，建议作者对第四次审稿人的意见做更加直接和具体的回答。

回应：感谢编委提出的宝贵意见。我们重新整理了思路，在提出 SEA 法的时候（见本文 2.2 部分）首先指出基于当前的认知诊断理论无法求出 KS 的标准误，只能采用变通的做法来处理。然后以结构方程模型为例，说明了这种变通作法是可以被接受的事实。之后，我们展示了边际分布和联合分布之间的关系，即根据属性边际概率可以获得 KS 后验概率的一个对应区间，在得不到 KS 标准误的基础上，至少能够为其提供一个精度的理论区间。需要指出的是，这个理论区间是最保守的估计，也就是说是最大的区间，加入其他的额外信息有可能缩小这个区间。另外，本文提出的 4 种新方法中，只有 SEA 法存在该问题。

希望我们的进一步修改能够让您满意。再次感谢您对本文的肯定！