国家医学考试中心

The National Medical Examination Center

临床执业医师分阶段实证研究临床思维能力测评系统的评分准确性研究

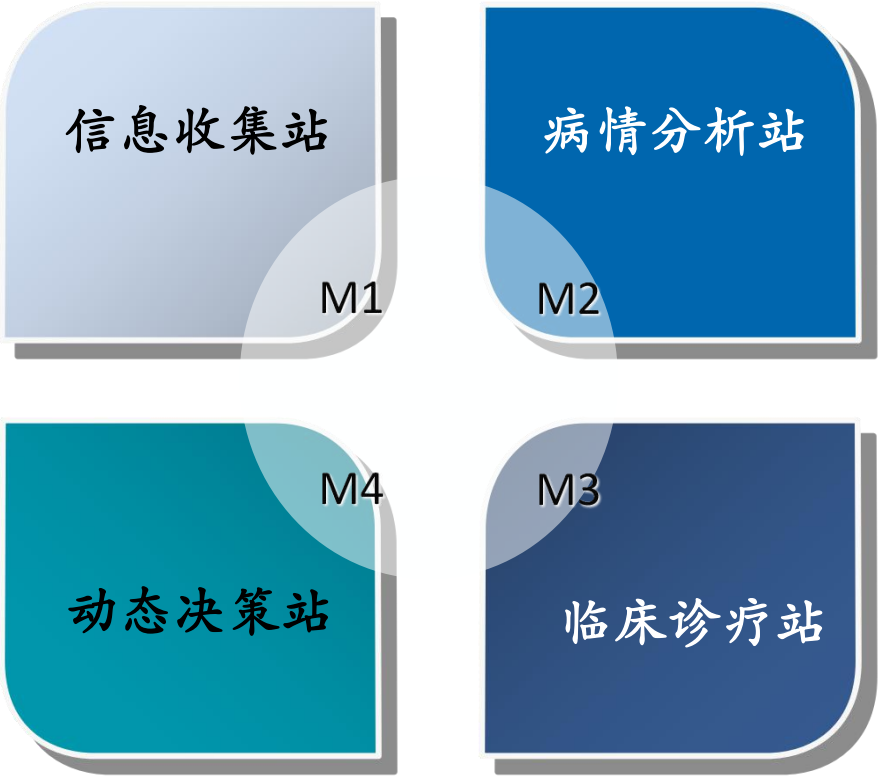
2019年10月22日星期二

国家医学考试中心
卢燕

目录 Contents

第一部分 研究背景

临床思维能力测评的结构



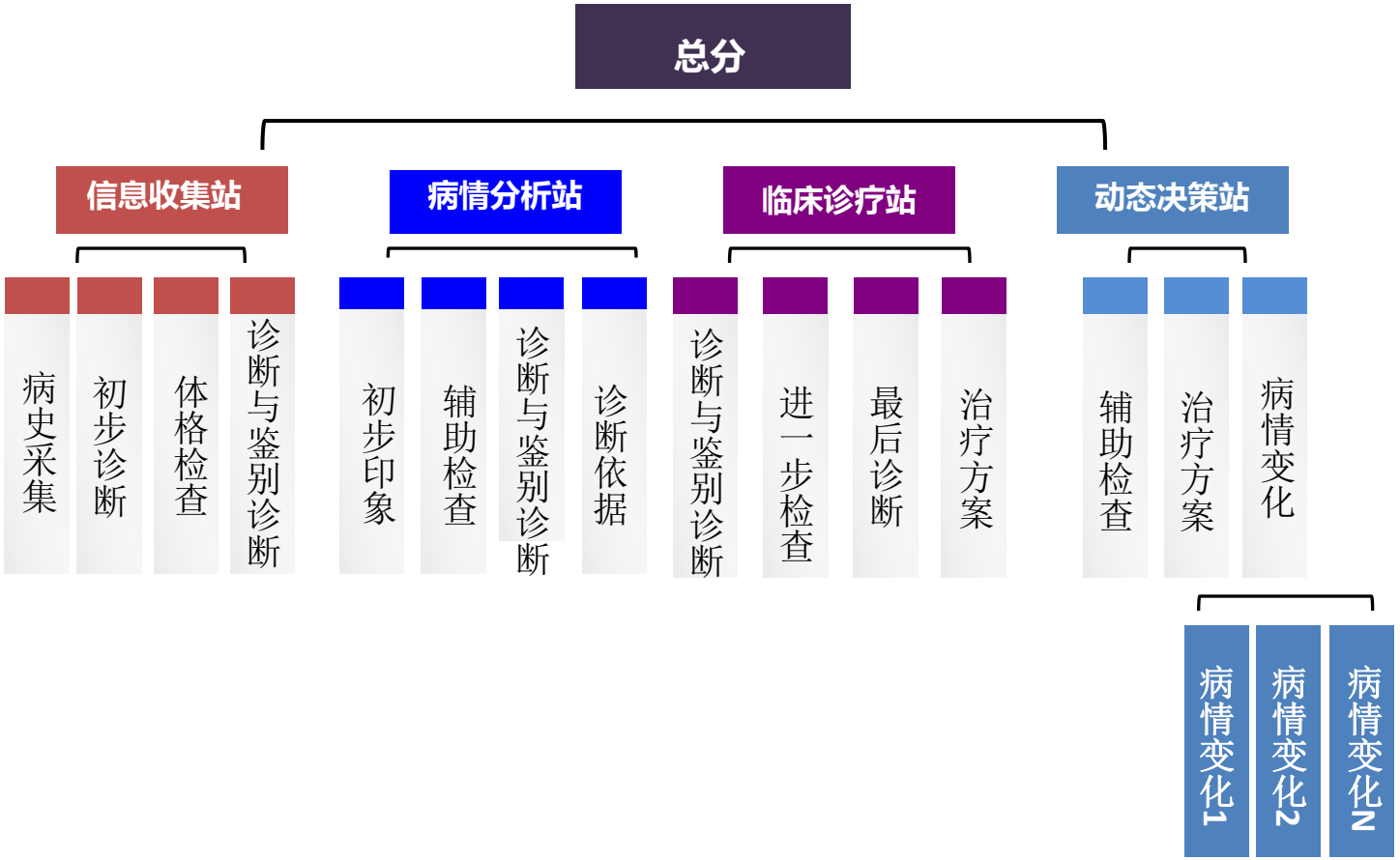
■根据设定的临床场景和患者就诊症状，完成病史采集并做出初步印象诊断；完成体格检查，并做出诊断与鉴别诊断。

■分析给定的患者信息和病历资料，做出初步印象诊断，提出明确诊断所必需的辅助检查项目，并根据检查结果做出诊断与鉴别诊断。

■根据给定的病历摘要内容，做出诊断与鉴别诊断，提出诊断和治疗所必需的进一步检查，根据检查结果做出最后诊断，并提出治疗方案。

■根据给定的临床场景和病历摘要内容，做出治疗，治疗后患者会发生新的病情变化，请根据病情进行必要的体格检查和辅助检查，并根据检查结果做出进一步治疗。

考站评分结构


总分
信息收集站
病史采集
初步诊断
体格检查
诊断与鉴别诊断
病情分析站
初步印象
辅助检查
诊断与鉴别诊断
诊断依据
临床诊疗站
诊断与鉴别诊断
进一步检查
最后诊断
治疗方案
动态决策站
辅助检查
治疗方案
病情变化
病情变化1
病情变化2
病情变化N

考站评分规则

命中

- 命中为给分项；
- 每个考点可根据重要性，设置“重要”或“次要”分组，不同组给分权重不同；
- 根据命中考点数占总答案数的比例，可设置得分比例。

顺序

- 顺序为扣分项，若顺序错将按权重扣分；
- 扣分有上限，且不会超过考生得分即不会出现负分；
- 病史采集模块及体格检查模块具有顺序评分。

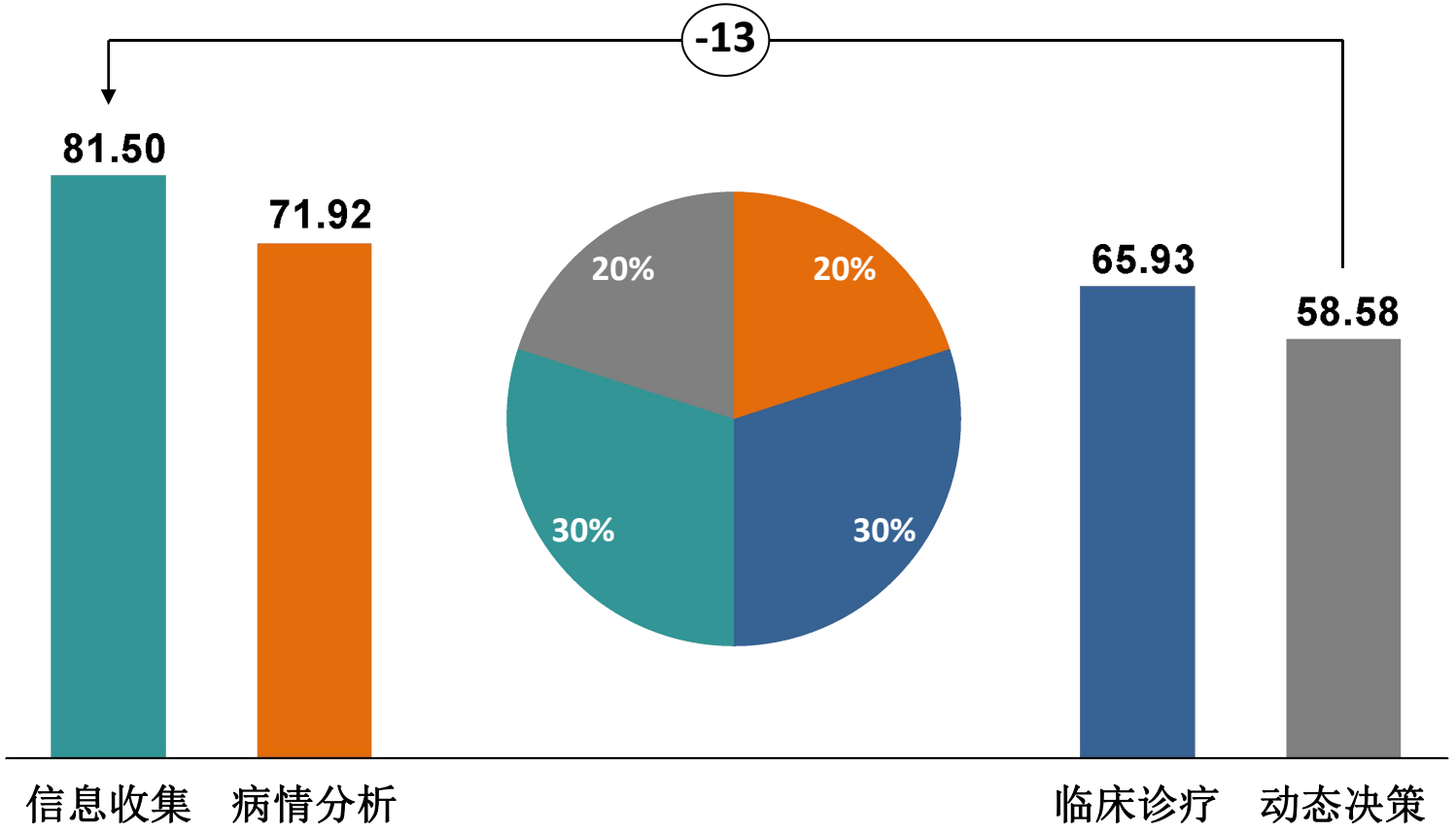
效率

- 效率为扣分项；
- 扣分有上限，扣分不会超过考生得分即不会出现负分；
- 每个模块均设置有答题数量上限，超过即为扣分；
- 可设置超出上限后扣分的权重，也可通过设置操作次数限制考生。

关联

- 可根据病史采集和初步诊断；体格检查和初步诊断；辅助检查和初步印象之间的关联正确命中占总关联命中数的比例设置权重；
- 奖励临床逻辑思维能力比较强的考生，无需命中所有信息，即可得到较高分数

临床思维能力测评各站掌握率



第二部分

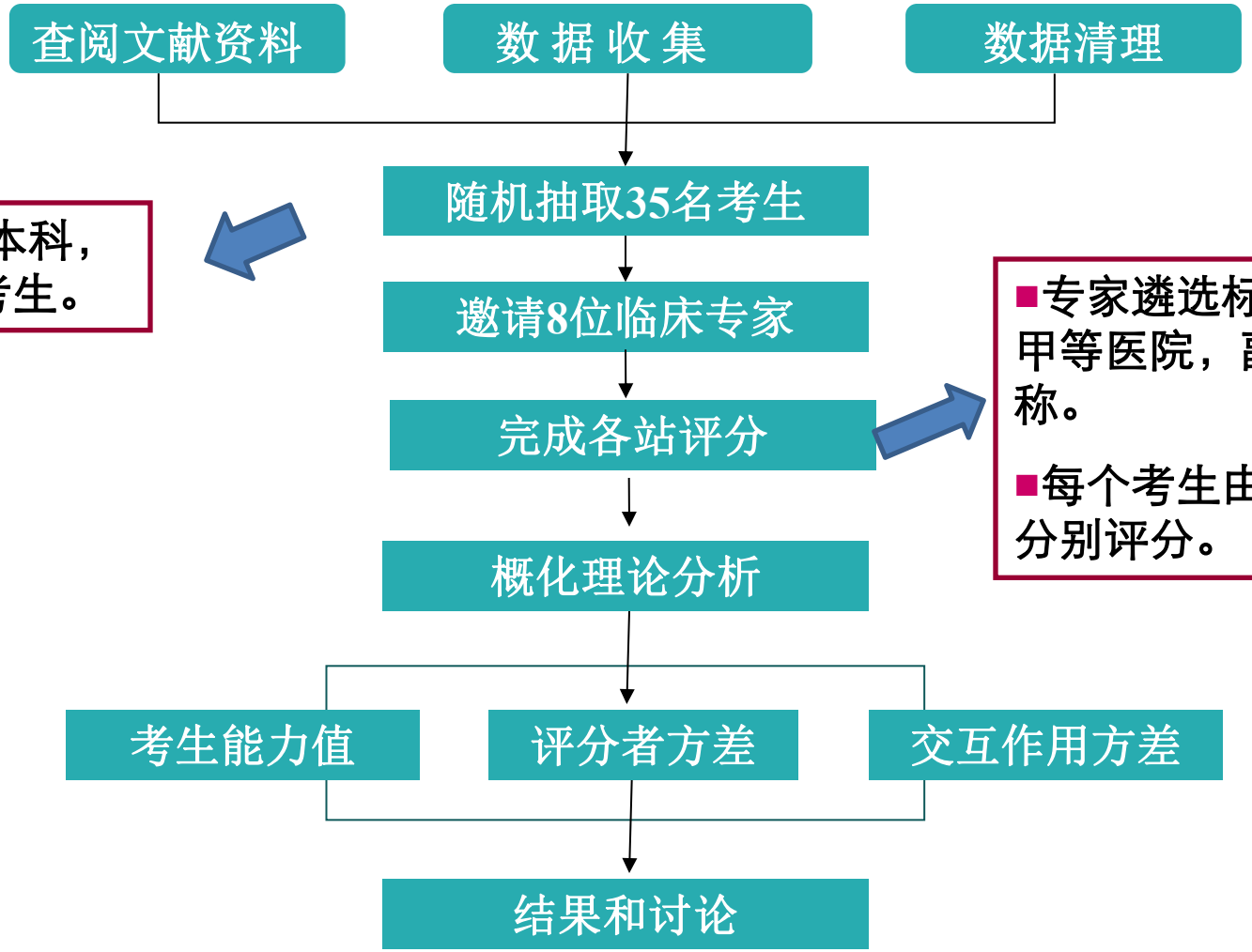
研究方法

本研究起因

- 计分规则复杂：包括命中（重要、次要等）；效率和顺序扣分；关联系数的计算……
- 规则的基础：命题专家根据自己的临床经验得出。

是否能真实反应考生的临床思维能力?

本研究流程



考生为5年制本科，2012年入学考生。

- 专家遴选标准为三级甲等医院，副高以上职称。
- 每个考生由两名考官分别评分。

概化理论- (Generalizability Theory , GT)

------GT理论，由克隆巴赫和格莱塞引进测量领域，为布伦南所大力发展。

------ GT理论核心—方差分析方法在测量学中的应用。

----GT理论解决问题—测评是否可靠？怎么能提高测评的可靠性。

关键词提取：

理论基础——经典测验理论，将测验各个侧面的方差进行分解。GT把观察分数的总体方差（分解成测量目标方差、侧面方差、各种交互作用方差，以及交互作用与其他不明的变异来源的混杂效应的残差方差部分。

研究基础——测量目标和侧面。 GT把测量者希望测量的那些实体称为测量目标（object of measurement）。GT用侧面（facet）这一概念来表示一组特定的测量条件，也即除测量目标外所有可能影响测量结果的因素。

研究步骤——概化研究和决策研究。

测量目标和侧面

GT把测量者希望测量的那些实体称为测量目标（object of measurement）。

GT用侧面（facet）这一概念来表示一组特定的测量条件，也即除测量目标外所有可能影响测量结果的因素。

本研究的测量目标及测量侧面

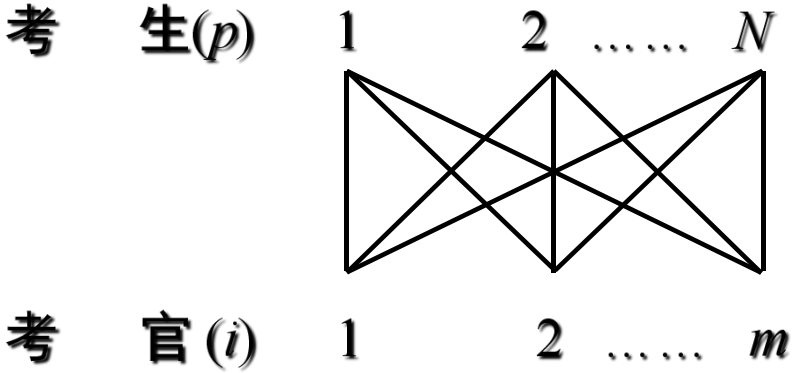
测量目标 → 参加2018年度CTA的考生在考试中表现出的临床思维能力

测量侧面

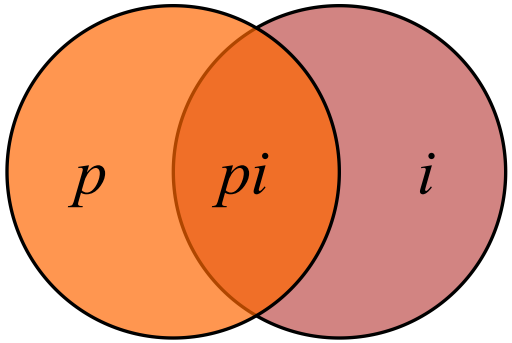
1. 题目，因为考生所考试题相同，本研究中视为相同，不包含
2. 考试情境，因为为标准化机考系统，本研究中视为相同，不包含
3. 评分者，本研究探讨的重点，将CTA系统评分视为一个评分者，将专家视为另外的评分者

研究期望，此部分方差在总体方差中占的比重越少越好，越少代表，CTA评分系统和考官评分之间的差异越小。

■研究设计及方差分解



交叉设计



考生主效应(p)

交互效应(pi)

考官主效应(i)

方差分解过程

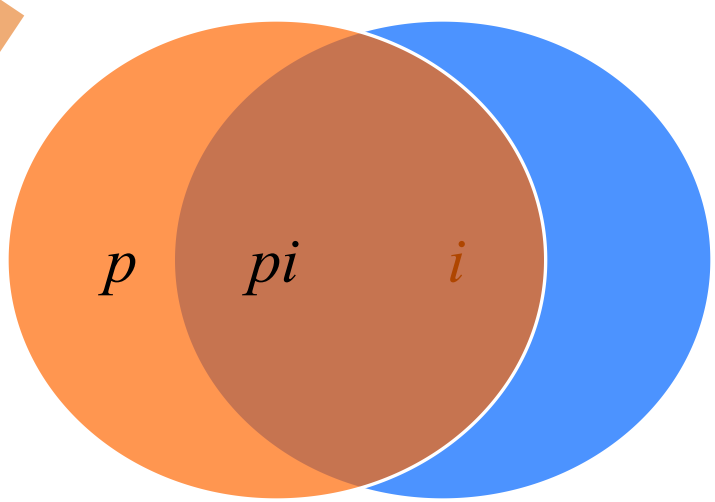
GT把观察分数的总体方差，分解成测量目标方差、侧面方差、各种交互作用方差，以及交互作用与其他不明的变异来源的混杂效应的残差方差部分。

总体方差=考生临床思维能力方差（测量目标p）+考官方差(r)+残差。

本研究的方差分解

在总体方差中的比例即为可靠性指标

在总体方差中占的比重越少，代表CTA评分系统和考官评分之间的差异越小。



考生主效应(p)

交互效应(pi)

考官主效应(i)

相对误差方差 和绝对误差方差

- 绝对误差方差指的是和测量主效应无关的所有其他效应引起的方差，具体到此研究，即和考生临床思维能力无关的方差。
- 相对误差方差指和测量主效应无关的所有交互效应引起的方差，具体到此研究，即考生能力和评分者的交互作用
- 绝对误差>相对误差

方差分解公式及可靠性系数

效应	方差和	自由度	均方差	变异分量的估计值
	SS	**df**	**MS**	
考生（p）	SSp	n_p-1	$MS_p=SS_p/df_P$	$\sigma^2_p=[MS_p-MS_{pi}]/n_i$
考官（i）	**SSi**	**n_i-1**	**$MS_i=SS_i/df_i$**	**$\sigma^2_{i=}[MS_i-MS_{pi}]/n_p$**
考生考官交互(p×i)	SSr	$(n_p-1)(n_i-1)$	$MS_{pi}=SS_{pi}/df_{pi}$	$\sigma^2_{pi}=MS_{pi}$

- 绝对误差——可靠性系数
  - 考官方差（i）+考官考生交互作用方差((p×i)
- 相对误差——概化系数
  - 考官考生交互作用方差((p×i)

第三部分 研究结果

■研究对象遴选标准

考生人数	考官人数	CTA
35	8	1

CTA系统考核和考生

■共四站，分别为信息收集、病情分析、临床诊疗、动态决策

■考生为5年制本科毕业，2012年入学考生

■专家遴选标准为三级甲等医院，副高以上职称

数据分析用软件

■概化理论分析

■使用该模型的配套软件

CTA系统与考官评分的相关

	信息收集	病情分析	临床诊疗	动态决策
相关系数	0.8237	0.6971	0.8559	0.8147
P值	0.0000	0.0000	0.0000	0.0000

- 为各考站专家评分与CTA系统评分的Pearson相关系数。
- 专家的评分和CTA系统评分的一致性很高，除病情分析站略低外，其他各站均达到了0.7以上；各站相关系数的双侧显著性检验结果都显著（$p<0.001$）。

■测量目标和测量侧面的方差分解

考站	考生数	p	比例(%)	r	比例(%)	p×r	比例(%)
信息收集	35	1.16	43.44	1.03	38.69	0.48	17.87
病情分析	35	1.28	67.72	0.20	10.80	0.41	21.48
临床诊疗	35	1.19	68.85	0.05	2.68	0.49	28.47
动态决策	34	1.91	56.75	0.66	19.63	0.79	23.61

- 对考生分数方差，从考生和评分者两个侧面进行的分解，上表为各侧面方差及方差所占总方差的比例。
- 除信息收集站达到了38.69%外，评分者侧面带来的方差占总方差的比例不足20%，尤其是诊疗决策站，评分者侧面带来的方差小于3% 。

可靠性研究结果

考站	考生数	全域分数	绝对误差	相对误差	概化系数	可靠性指数φ
信息收集	35	1.1566	0.5019	0.1586	0.8794	0.6736
病情分析	35	1.2843	0.2040	0.1358	0.9044	0.8629
临床诊疗	35	1.1829	0.1784	0.1630	0.8789	0.8689
动态决策	35	1.9086	0.4849	0.2648	0.8782	0.7974

- 绝对误差方差指的是和测量主效应无关的所有其他效应引起的方差，具体到此研究，即和考生临床思维能力无关的方差；
- 相对误差方差指和测量主效应无关的所有交互效应引起的方差，具体到此研究，即考生能力和评分者的交互作用。
- 由表可知，相对全域分数方差，相对误差在各站所占的比例都较小。各站的概化系数和可靠性指数φ均较高。
- 说明评分者组内，无论评分者是专家还是CTA系统得分，一致性都很好。

- 此结果可知，考生的临床思维能力分数的差异，绝大部分是考生自身能力差异造成的，评分者侧面，无论是专家还是系统评分带来的差异都很小
- 结果说明，2位专家与CTA系统评分对考生评价的差异很小，评分标准趋于一致。

第四部分

下一步研究

2019年评分一致性研究

------继续2018年的研究，组织专家对考生作答评价；

------解决考生样本量小的问题，进一步加大样本量；

------对一致性较低的考站，进行深入研究，分析原因。

关键词提取：

多面Rasch模型——（The many-facet Rasch model，MFRM）。是Linacre在丹麦数学家Rasch开发的单参数Rasch模型的基础上，将评分者评价的宽严度作为一个参数加入，从而构成了被试的能力、项目的难度和评分者宽严度的多参数模型。

$$\log\left[\frac{P_{nijk}}{P_{nij(k-1)}}\right] = B_n - D_i - C_j - F_k$$

对能力维度的结构验证

因子分析

------通过研究众多变量之间的内部依存关系，探求观测数据中的基本结构，使用少数潜在因子构建模型。

------ 主要分为：探索性因素分析和验证性因素分析。

----测验评估领域，主要用于降维，探索和验证测验所考核的能力结构是否符合理论预期。

关键词提取：

探索性因子分析——（Exploratory Factor Analysis，EFA），通过对变量或样本的相关系数矩阵内部结构研究，利用降维的思想，找出少数几个影响所有数据的潜在因子。

验证性因子分析——（Confirmatory Factor Analysis，CAF），利用结构方程（Structural Equation Models，SEM）的理论，结合了路径分析和因子分析，对构建的潜在变量和观察变量之间的模型结构，进行验证。

国家医学考试中心

National Medical Examination Center

谢　谢