科学通报 2026 年 第 71 卷 第 11 期: 2472 ~ 2484
评 述
《中国科学》杂志社
SCIENCE CHINA PRESS

认知任务的测量信度: 进展与前景

朱芃芃[1,2†], 刘铮[3†], 康春花[4*], 胡传鹏[1,2*]

1. 南京师范大学心理学院, 南京 210097
2. 江苏省高校哲学社会科学实验室, 南京师范大学青少年教育与智能支持实验室, 南京 210097
3. 香港中文大学(深圳)人文社科学院, 深圳 518172
4. 浙江师范大学浙江省儿童青少年心理健康与危机干预智能实验室, 金华 321004

† 同等贡献

* 联系人, E-mail: hcp4715@hotmail.com; akang@zjnu.cn

2025-04-18 收稿, 2025-06-11 修回, 2025-07-18 接受, 2025-07-23 网络版发表
国家自然科学基金(32471097)和浙江省儿童青少年心理健康与危机干预智能实验室重点开放基金(23MHCICAZD04)资助

摘要 认知任务(cognitive tasks)是研究人类认知过程的核心手段, 广泛应用于认知科学与神经科学等领域. 随着个体化研究的兴起, 认知任务在测量个体差异方面的应用日益增多, 其测量信度问题也逐渐引起关注. 近年研究发现, 一些在群体层面呈现稳定实验效应的任务, 在个体层面的信度却表现不佳, 形成所谓的"信度悖论". 深入分析表明, 该问题主要源于两方面挑战: 其一, 构念效度不足, 任务指标未能有效反映个体在潜在认知能力或过程上的差异; 其二, 传统信度估计方法难以适应认知任务所呈现的层级结构数据. 前者强调需提升任务指标对潜在认知能力或过程的测量效度, 后者则表明亟需发展更契合数据结构的信度评估方法. 近年来, 研究者使用基于置换检验的分半信度、内相关系数(ICC)等方法作为认知任务的信度估计方法, 但关于如何选取能稳定反映潜在认知能力或过程的指标, 仍有待深入探索. 要提升认知任务的信度, 仍需在构念效度提升、测量误差控制、统计建模优化及测量模型创新等方面开展系统研究.

关键词 认知任务, 信度悖论, 信度, 个体差异, 被试间差异

认知能力是人类智能的核心组成部分, 涵盖知觉、注意、记忆、执行功能、决策及语言处理等多个方面[1]. 这些基本认知能力不仅构成人类行为和思维的基础, 还直接影响学习、工作、社会交往乃至心理健康的多种结果[2]. 因此, 深入理解认知能力及其个体差异, 对于揭示人类认知机制、促进个性化干预策略的制定, 具有重要的理论与实践意义.

认知任务(cognitive task)作为实验心理学和认知科学中的关键研究工具, 近年来也被用于测量个体在执行不同认知活动中的表现. 自20世纪以来, 实验心理学通过严格控制的认知任务探讨认知机制, 这些任务范式也在神经科学、精神病学和人工智能等领域得到了广泛应用[3]. 近年来, 得益于计算建模(computational modelling)和神经影像技术的迅猛发展, 认知任务不仅成为研究认知过程的核心方法, 还在揭示潜在认知能力或过程的个体发育轨迹及其临床诊断价值中发挥了重要作用[4,5]. 例如在精神病学研究领域, 认知任务与计算建模的结合显著提升了对认知异常及个体差异的理解[6~8], 并推动了个体化诊疗策略的发展[9,10].

信度作为心理测量学中的核心概念, 起源于经典测量理论(classical test theory, CTT). 该理论认为观测分数由真实分数与误差分数两部分组成, 其中真实分数反映个体的稳定特征, 而误差分数代表随机波动[11,12]. 信度衡量的是测量工具在不同时间和情境下

引用格式: 朱芃芃, 刘铮, 康春花, 等. 认知任务的测量信度: 进展与前景. 科学通报, 2026, 71: 2472–2484
Zhu P, Liu Z, Kang C, et al. Measurement reliability of cognitive tasks: current trends and future directions (in Chinese). Chin Sci Bull, 2026, 71: 2472–2484, doi: 10.1360/CSB-2025-0551

© 2025《中国科学》杂志社

对同一特征进行重复测量时所得结果的一致性与稳定性[13]. 在以往的基础研究与应用实践中, 认知任务多用于实验研究, 用于揭示群体在不同实验操作下的差异从而帮助理解人类的认知规律. 然而, 当认知任务用于个体差异测量时, 其信度问题却长期被忽视[14~16]. 早在2003年, Vasey等人[17]就指出计算精神病学领域对任务范式信度关注不足的问题; 然而直至2019年, Parsons等人[15]仍发现认知任务信度检验的实践相对稀少. 例如, 尽管点探测范式(dot-probe paradigm)被广泛采用, 但在数千篇已发表文献中, 仅有13篇报告了该任务的信度指标.

近年来, 心理学领域面临的可重复性危机[18~20]进一步促使研究者重审认知任务在个体水平上的信度表现(如文献[15,21~23]). 相关研究发现, 多数经典认知任务在测量个体差异时的信度表现不佳(表1), 引发了对其作为个体化测量工具适用性的关注.

这些研究表明, 尽管多数经典认知任务在群体层面可获得稳定的实验效应, 其测量信度却往往较低, 这一现象被称为“信度悖论”(reliability paradox)[23]. 认知任务的低信度不仅会削弱研究结论的可靠性和可重复性, 还可能误导对个体认知能力的评估, 影响临床决策和干预效果. 因此, 深入理解和系统评估认知任务的信度现状, 并积极探索提升信度的有效策略, 已成为当前亟需解决的重要议题.

尽管近年来学界对认知任务信度的关注日益增加, 但系统性的理论分析与方法讨论仍较为缺乏. 本文将首先分析信度悖论产生的可能机制与测量挑战, 继而系统介绍适用于认知任务的多种信度计算方法, 最后探讨影响认知任务信度的关键因素与提升策略, 以期为认知科学、计算精神病学等领域的个体测量研究提供参考.

1 认知任务的信度悖论及其成因

“信度悖论”的成因可追溯至心理学科发展早期实验取向与问卷测量取向的分化[26]. 实验研究侧重群体水平的效应显著性与可重复性, 而问卷研究则强调个体层面测量的稳定性与区分力[17]. 在实践中, 实验取向更关注实验操作的有效性, 即被试在不同条件下作为整体所表现出的平均差异; 而问卷取向则强调如何测量个体差异, 进而发展出心理测量学中诸多关键理论、方法与工具. 当研究者将原本用于检验实验操纵效应的认知任务用于个体差异测量时, 可能出现“信度悖论”.

在实践中, 认知任务多以验证实验操控效应为目的, 强调内部效度(internal validity), 即多大程度上能将因变量的变化归因于自变量操控而非混淆变量[27]. 由于单个试次的实验数据受到随机因素(如注意力波动、环境干扰等)的影响, 实验者常通过标准化实验环境、标准化流程和对多个试次数据取均值等方法消除随机噪音的干扰, 获得稳健的群体效应. 然而, 这种强调内部效度的设计不仅压缩了个体间变异, 也形成了与问卷研究不同的数据结构. 当研究目标从验证因果效应

表 1 部分常见认知任务信度的主要结果

Table 1 Key findings on the reliability of common cognitive tasks

源文献	作者	主要任务	信度类型	主要结论
What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis[21]	Elliott等人	常见fMRI任务(如工作记忆、情绪加工、决策)	重测信度(ICC)	基于90项实验、1008名被试的元分析结果表明, 用于fMRI研究的认知任务的平均信度较低(ICC=0.397), 在Human Connectome Project和Dunedin样本中, 11项任务的ICC介于0.067~0.485之间.
Large-scale analysis of test–retest reliabilities of self-regulation measures[22]	Enkavi等人	自我调节相关的行为任务 & 量表	重测信度(ICC)	基于154篇文章和17550人的数据, 36项认知任务的重测信度较低(均值0.610), 且信度随样本量增加而下降
The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences[23]	Hedge等人	7种经典认知任务(flanker、stroop、stop-signal、go/no-go、posner cueing、Navon、SNARC)	重测信度(ICC)	7种认知任务的ICC范围为0~0.82, 多数低于可接受水平
Impulsivity is a stable, measurable, and predictive psychological trait[24]	Huang等人	与冲动性相关的自评量表和行为任务(48种冲动性测量, 来源于10个自评量表和10个行为任务)	重测信度(ICC)	基于1676人与198人(重测)的数据, 来源于10个认知任务的20个实验效应指标的ICC中位数为0.44
Individual differences in computational psychiatry: A review of current challenges[25]	Karvelis等人	计算精神病学中的多种任务(如赌博任务, 联结学习任务)	Pearson相关系数, 重测信度(ICC)	回顾20篇研究发现, 来自20个认知任务的计算模型参数信度较低(相关系数范围为r=0.07~0.91, ICC范围为0.25~0.91)

转向测量个体差异时，这些设计限制了个体差异的展现, 降低了任务在测量个体差异方面的信度[23].

对内部效度的追求也在一定程度上忽视了对构念效度(construct validity)的系统检验. 认知任务结果指标的多样性是这种忽视的表现. 构念效度指某一测量工具是否能够合理、有效地反映其所声称的心理构念(psychological construct)[27]. 在认知任务的背景下, 构念可以指无法直接观测的认知能力或过程[28]. 若一个工具被用于测量个体在特定认知能力或过程上的差异, 其测量指标应能够稳定地反映该构念所指向的能力或过程[29,30]. 然而, 多数认知任务源于实验研究传统, 其设计优先关注内部效度, 而非构念效度. 因此, 即使关键指标(如反应时差、错误率)在条件间存在显著差异, 也未必能有效反映潜在的心理构念. 当此类尚未经过构念效度验证的指标被用于测量个体差异时, 其结果在多大程度上能够测量出研究者关心的潜在认知能力或过程往往是未知的, 最终表现出低信度也不意外. 这可能在一定程度上解释了为何许多理论上与特定心理能力相关的认知任务, 在实际应用中难以稳定地测量个体差异.

综上所述, 信度悖论揭示了实验研究与问卷测量在研究目标与方法论上的长期分野对当前跨领域应用所带来的深远影响. 近年来, 研究者围绕信度悖论进行了大量研究, 尤其是针对认知任务数据结构的特点进行了信度估计方法的开发与测试, 同时据此对诸多经典认知任务的信度表现进行了系统评估(表1). 然而, 针对更为本质的构念效度问题, 相关研究仍较为有限.

2 认知任务数据结构的挑战及其解决方案

对认知任务的信度评估面临的直接挑战是层级结构的数据. 认知任务往往涉及多种实验条件, 每个被试在每个条件下均需要完成多个试次, 形成了层级化的数据结构(如图1左侧所示, 以自我匹配任务[31]为例). 这种数据结构与研究者寻找实验条件间差异的目标相适应, 但与传统问卷测量中“每个被试在每个条目上仅有一个得分”的情境不同, 所以需要适配相应地信度评估方法.

针对上述问题, 研究者提出了两类适用于认知任务数据结构的信度评估指标[33,34]: 一是分半信度, 用于衡量个体在任务执行过程中不同试次间的一致性; 二是重测信度, 以评估个体在不同时间点上的任务表现稳定性.

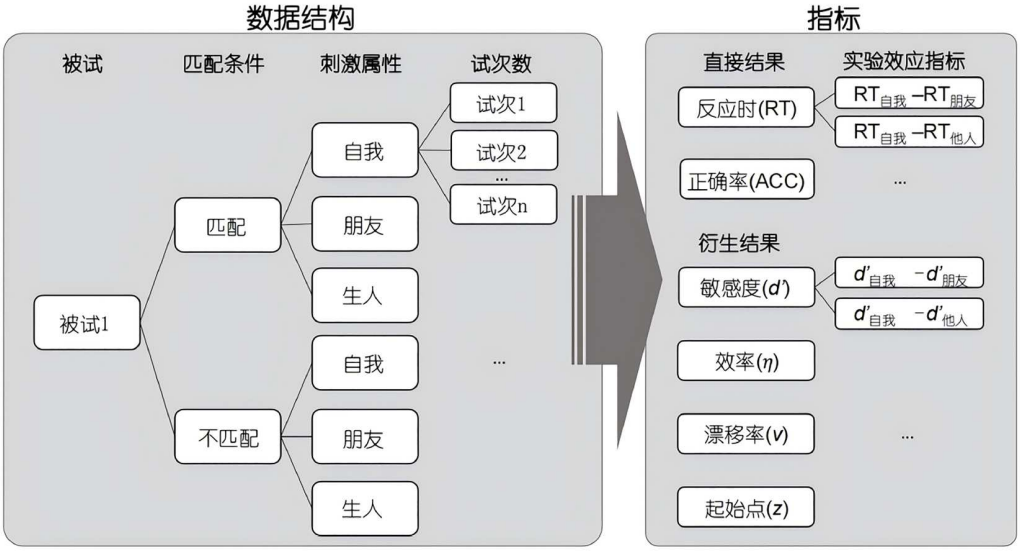


图 1 自我匹配任务的数据结构及指标. 该任务的衍生结果包括基于信号检测理论的敏感度指标(d')、效率指标衡量反应时与正确率的权衡(计算公式: RT/ACC)以及基于扩散漂移模型(drift-diffusion model, DDM)的参数漂移率(v)和起始点(z), 后两者分别表征证据积累的平均速率与决策边界的起始位置(详见文献[32])

Figure 1 Data structure and potential dependent variables of the self-matching task. The derived indices include sensitivity index (d'), which is based on signal detection theory; the efficiency index, which is calculated by RT/ACC; the drift rate (v) and starting point (z), which are key parameters of the Drift-diffusion model and representing the rate of evidence accumulation and the initial decision boundary position, respectively (see Ref. [32] for details)

2.1 认知任务的分半信度

认知任务结果数据的结构使得传统内部一致性指标(如Cronbach's α)难以适用[15,33]. 一方面, α的计算依赖测量条目对真实得分贡献相等(即tau等值假设)[35], 而认知任务中各试次往往具有异质性与条件依赖性, 该假设难以满足; 另一方面, Cronbach's α需对每位被试的试次反应取平均生成"子得分", 此做法忽略了试次层级的变异性, 且在试次数不等时赋予等权重, 可能放大不稳定得分的影响. 此外, 该方法排除了试次误差, 仅衡量子得分间一致性, 无法反映个体指标的真实稳定性. 更重要的是, α假设被试对所有"条目"反应一致, 这一统计前提也常与认知任务的理论基础不符[33]. 因此, α不仅在统计上存在偏误风险, 也在方法论层面与认知任务之间存在错配. 鉴于上述限制, 近年来研究者逐渐转向更适合认知任务数据结构的信度评估方法, 例如分半信度.

具体而言, 针对认知任务的多条件、多试次特点, 研究者需选择合适的分半方法, 以确保信度估计的稳健性. Pronk等人[34]系统评估了四种常用的分半方法(表2), 结果表明置换分半法(permutation split-half method)最为稳健. 该方法通过多次无放回随机抽样, 将实验试次分为两半, 并计算每次分半后Spearman-Brown矫正的皮尔逊相关系数, 从而得到信度系数分布. 这一方法能够有效减少单次分半可能带来的抽样误差, 同时综合考虑实验设计中的时间效应、任务效应及非线性变换的影响, 为信度估计提供更为稳健的测量依据[34]. 后续研究发现, 该方法能够广泛应用于多种认知任务[36~38].

目前, 以上的多种分半计算信度的方法均可通过R统计软件中的*splithalfr*包实现[39]. 网页(https://www.scidb.cn/en/anonymous/QVJqbWFt)提供了以自我匹配任务的反应时数据为例[37], 计算前后分半、奇偶分半及置换分半信度的示例代码.

2.2 认知任务的重测信度

重测信度(test-retest reliability)主要用于评估同一测量工具在不同时间点对同一被试重复测量结果的一致性. 早期研究通常采用皮尔逊相关系数、配对样本*t*检验和Bland-Altman图等方法评估重测信度[40], 然而这些方法均未能全面衡量重测信度的核心特征[41]. 相较之下, 内类相关系数(intraclass correlation coefficient, ICC)能够同时反映测量结果的相关性与一致性, 近年来被广泛使用.

在实践中, ICC计算方法多样[42]. 在认知任务的重测信度计算中, 研究者建议采用同时报告双向随机效应模型ICC (2,1)和双向混合效应模型ICC (3,1)[15,41]. ICC (2,1)(双向随机效应模型)适合对结论进行推广的目标; ICC (3,1)(双向混合效应模型)更适合研究目标群体为特定被试群体的研究. 相关ICC计算公式如下:

$$\mathrm{ICC2}=\frac{\mathrm{MS_{BS}}-\mathrm{MS_E}}{\mathrm{MS_{BS}}+(k-1)\mathrm{MS_E}+\frac{k}{n}(\mathrm{MS_{BM}}-\mathrm{MS_E})}, \quad (1)$$

$$\mathrm{ICC3}=\frac{\mathrm{MS_{BS}}-\mathrm{MS_E}}{\mathrm{MS_{BS}}+(k-1)\mathrm{MS_E}}. \quad (2)$$

其中, $\mathrm{MS_{BS}}$是被试间方差(between-subject mean square), 反映个体间存在的差异; $\mathrm{MS_E}$是误差方差(error mean square), 反映随机误差或噪声; $\mathrm{MS_{BM}}$是测量间方

表 2 不同分半方法的优点、缺点及适用条件

Table 2 Comparison of different split-half methods: advantages, limitations, and applicable conditions

分半方法	描述	混淆效应				优点	缺点	适用性
		时间效应	任务设计效应	试次抽样效应	非线性评分			
前后分半(first–second methods)	根据试次序号分为前后两半	×	√	×	×	简单易实现	与时间效应(如参与者的疲劳或学习效应)混淆	当时间效应不显著或可控时可使用, 但如果任务可能涉及学习或疲劳效应, 应避免使用
奇偶分半(odd–even methods)	根据试次序号分为奇偶两半	√	×	×	×	有效控制时间效应	与任务设计(如任务包含交替的条件)混淆	任务设计中如果存在交替条件时, 需要谨慎使用
置换分半(permutated methods)	无放回的随机分半	√	√	√	√	多次抽样, 结果稳健	几乎没有	适用于通过多次抽样来获得稳定估计的情况
蒙特卡罗分半(Monte Carlo methods)	有放回的随机分半	√	√	√	√	不基于任务得分和试次之间存在线性关系的前提假设	当试次数少时, 重复抽样导致结果方差降低, 从而可能高估信度[33]	适用于采用非线性评分的任务, 但对于试次数较少的实验设计需谨慎

差, 通常反映重测任务间的变异性; k是试次; n是被试总数. ICC (2,1)和ICC (3,1)的计算可通过R语言*psych*包实现[43], 详情见网页链接https://www.scidb.cn/en/anonymous/QVJqbWFt.

在使用ICC评估信度时, 所需的样本量也是一个值得关注的问题, 尽管目前尚无统一标准. 但根据经验法则, 建议样本量至少达到30[41]. ICC解释通常遵循以下标准: ICC<0.5表示信度较低, 0.5~0.75为中等水平, 0.75~0.9为良好, >0.90则可认为具有较高的信度[44,45]. 然而, 这些标准仅作为一般性参考, 具体解释需结合研究背景、测量工具及数据特性加以判断.

3 认知任务实验效应指标的多样性挑战

认知任务的信度研究中另一个挑战是其指标的多样性. 以Stroop任务为例, 常用数据包括反应时(reaction times, RT)和错误率(error rate), 且可构建多个指标, 如一致/不一致条件下的反应时与错误率、两条件之差值, 以及更复杂的衍生指标[25], 如将反应时与错误率结合的效率指标(efficiency), 或依据认知建模方法估计的潜在认知参数, 如证据积累模型中的信息累积速率(drift rate, v)等. 但如前所述, 这种指标多样性反映了认知任务在构念效度上的薄弱: 不同指标可能涉及不同认知过程, 其是否指向统一的构念, 缺乏系统性验证.

构念效度的不明确性对信度评估带来的影响, 已在多项实证研究中得到体现. 大量实证研究显示, 即便在同一任务中, 不同指标的信度差异显著. Hedge等人[23]对七类经典认知任务分析发现, 各任务内部指标的信度范围广泛(0~0.82). 例如, Stroop任务中一致条件下反应时信度达0.77, 而错误率仅为0.36. 类似地, Liu等人[37]在自我匹配任务中也发现, 反应时和效率指标在内部一致性和重测信度方面优于其他指标(如正确率、敏感度、扩散模型参数). 然而, 反应时的优势并非普遍规律. von Bastian等人[46]系统分析了注意控制领域不同类别(如抑制、转换、心智游离)的任务及其指标(反应时、准确性、敏感性等)的信度, 未发现一致性规律. 该研究表明, 认知任务信度随任务类型显著变化. 例如, 心智游离(mind wandering)类认知任务的信度中位数较高(0.90; M=0.89, SD=0.08), 而抑制(inhibition)类任务的信度中位数相对较低(0.79; M=0.72, SD=0.22), 且抑制类任务的信度系数分布范围更大. 这些现象表明, 许多任务缺乏明确的构念效度支持. 不同指标若对应不同心理机制, 其信度差异实属必然. 指标层面的构念模糊性, 是造成信度分化的重要来源.

此外, 用于测量相同认知过程的任务间存在高度碎片化, 亦反映构念效度的不足[30]. 这一问题体现在两个方面: 一方面, 同名任务在实施上差异显著, 如自我参照编码任务在流程设计上高度异质[47]; 另一方面, 测量同一构念的不同任务相关性普遍偏低. 例如, 在测量延迟满足时, 延迟折扣任务、自我控制问卷与棉花糖实验等指标间常无显著相关性[22,24,48]. 这一“低任务间一致性”现象可能源于构念操作定义不同、行为指标敏感性差异或控制混淆变量能力不足. 结果是, 虽然任务声称测量相同心理能力, 实际却可能映射到不同的成分, 构成典型的混杂谬论(jingle-jangle fallacies)[49]. 该碎片化不仅阻碍构念的一致建构, 也削弱了测量信度与效度的基础.

4 提高认知任务信度的途径

认知任务的信度已成为认知科学和心理测量领域的共同关注点[50]. 前文分析表明, 认知任务信度偏低的根本原因在于, 实验研究的设计目标通常不同于问卷研究, 这导致认知任务设计上未充分考虑其在测量个体差异方面的表现. 尽管已有研究者根据认知任务数据的层级结构提出了相应地信度估计方法(见第3小节), 信度提升仍需多角度、全方位的测量学考量. 近年来, 一些研究已从个体差异与统计层面开展了相关工作(如[12,51~60]), 但一个相对较少被关注且亟待解决的问题, 是前文反复提及的“构念效度”本身. 提升认知任务的构念效度, 即确保任务及其指标能够准确、一致地测量其所声称的心理构念, 这一直是认知测量领域的一个薄弱环节, 相关的系统探讨和实践相对匮乏. 相较于前人侧重于优化测量过程本身(如减少测量误差、改进模型), 本文指出, 回答认知任务到底“测量了什么”这一根本问题更为重要. 因此, 本文将提升认知任务信度的策略分为四类(表3), 涵盖提升任务构念效度、增强个体间变异、降低测量误差、发展更稳健的信度估计模型以及拓展传统测量框架.

4.1 提升认知任务对潜在认知过程或能力的捕捉能力

要解决认知任务指标可能无法有效捕捉并稳定反映所欲测量的潜在心理构念这一问题, 研究者需要着力提升任务本身的构念效度. 研究者可以借鉴心理测量学与认知科建模的方法, 从两个方向对任务进行“构

表 3 提高认知任务信度的主要策略

Table 3 Strategies for improving reliability of cognitive tasks

策略目标	具体策略	策略描述	相关研究
提升构念效度	多任务整合建模	整合多种任务结果, 构建合成指标, 采用统计建模方法提取稳定潜变量	[24,61]
	计算建模	使用计算模型参数揭示个体差异, 提高信度, 并增强任务指标的心理学解释力	[62,63]
增强个体间变异、降低测量误差	调整任务难度	调整任务难度, 避免天花板/地板效应	[51,52]
	加入游戏化设计	通过融入角色扮演、互动反馈机制等元素, 增加任务复杂性	[53~55]
	增加样本多样性	采用更广泛的被试来源, 避免单一群体	[51,56,57]
	增加正式实验试次数量	通过增加试次减少随机误差	[37,64]
	增加练习	通过充分练习帮助被试达到稳定的表现	[64,65]
	控制实验环境	确保被试注意力集中, 减少外部干扰	[66]
发展更稳健的信度估计模型	使用分层模型	使用分层建模方法, 提高估计稳定性并提升总体信度	[58~60]
	使用新型信度指标	应用SNR、dblICC等新指标, 减少模型依赖, 提升评估准确性	[12,67]
拓展传统测量框架	引入现代测量理论	引入现代测量理论方法, 增强对试次变异、任务维度、情境变化等误差来源的建模能力	[53,68,69]

念重建”: 一是通过多任务整合, 运用潜变量建模(如因子分析、结构方程模型)提取更稳定、具理论指涉的合成指标; 二是通过认知建模, 从单一任务的行为数据中估计具体认知过程的模型参数.

多任务整合方面, Heaton等人[61]基于NIH Toolbox Cognition Battery (NIHTB-CB)中的7个评估5类核心认知功能的子任务, 计算出晶体智力、流体智力与总体认知能力三类合成得分. 实证研究表明, 这些合成指标具有较高的内部一致性(Cronbach's α=0.77~0.84)与重测信度(r=0.86~0.92), 同时在收敛效度和区分效度方面表现良好, 与个体的教育背景、健康状况和社会功能等变量密切相关. 类似的, 在人格测量领域, Huang等人[24]基于对1676名被试进行的48项冲动性相关测量, 采用双因素模型(bifactor model)提取出一个融合问卷与认知任务指标的通用潜在因子“I”. 结果显示, 该通用因子在重测中表现出高度稳定性(r=0.85), 并在预测现实生活中的冲动行为方面优于现有单一问卷或任务指标. 然而, 这种“多任务整合+潜变量建模”的方式依赖于一个重要前提: 所整合的任务确实测量某一统一的心理构念. 但这一前提在常常难以满足, 许多表面上测量相同认知能力(如执行功能、冲动控制或工作记忆)的任务, 在实证中却常表现出较低的相关性[22,24,48], 呈现出“构念碎片化”特征. 在这种“构念碎片化”的情况下, 难以提取稳定的因子, 其信度水平也因此受到限制.

认知建模是另一个可能的路径, 即试图从个体在单一任务中的行为数据中提取更贴近心理机制的模型参数. 这些模型参数(如证据积累模型中的漂移率、决策边界)通常被认为与具体认知过程高度对应, 理论上更贴近个体差异的心理基础. 研究表明, 计算模型参数相较于传统行为指标, 在特定情境下具有更高信度[62,63,70]. 例如, Rappaport等人[70]在Eriksen flanker任务中, 发现扩散漂移模型参数(决策阈值、漂移率、非决策时间)比反应时和正确率等传统指标展现出更高的信度和一致性, 且能有效区分决策差异. 然而, 也有研究发现, 模型参数信度并非总是高于直接任务结果, 有时甚至被评为中等或更低[37,71]. 例如, Liu等人[37]在自我匹配任务中评估了标准扩散漂移模型在自我匹配任务中的适用性, 发现漂移率(v)和起始点(z)的分半信度和重测信度均低于可接受水平. 模型参数信度不足主要源于模型复杂度带来的共线性及过拟合问题[22,72], 以及模型与认知过程的匹配度不足[32,73].

无论是多任务整合中的潜变量建模, 还是基于行为数据的计算模型构建, 研究者都需要考虑如下问题: 用来测量个体差异的指标是否能够所期待测量的心理过程. 唯有在模型建构过程中始终关注变量与构念之间的关系, 强化认知机制的建模基础, 才能实现认知任务信度的提升.

4.2 在任务设计与实施过程中的信度提升

认知任务在追求内部效度的过程中, 往往因标准化设计和试次噪音而压缩了个体差异, 并形成了不利

于稳定测量个体差异的数据结构. 因此, 在任务设计与实施阶段进行优化仍有提升信度的空间. 根据经典测量理论, 测量信度取决于被试间的真实变异相对于测量误差的比例. 在这一理论视角下的关键挑战在于, 认知任务中个体间差异往往被大量的试次层级噪声所掩盖. Rouder等人[73]对Flanker任务与Stroop任务的数据重新分析发现, 试次层级的随机误差——即在同一被试和相同实验条件下, 不同试次间反应时间的波动——大约是被试间真实差异的8倍. 这种高水平的试次噪音主导了总体变异结构, 显著削弱了个体差异信号的表达, 导致传统相关性分析难以稳定捕捉个体间的共变关系, 从而严重影响任务指标的信度估计. 在此背景下, 优化认知任务信度的可行策略包括: 提升被试间差异的可测性(如通过更有区分度的任务设置、引入更异质的样本), 以及减少测量误差的干扰(如增加证实实验的试次数量、增加练习、控制实验环境等).

4.2.1 增加被试间差异

任务标准化而导致的个体间差异被压缩, 需要在任务设计和样本选择上采取措施, 以增加被试间差异. 提升被试间表现差异可通过调整任务难度、引入游戏化设计以及扩大样本的多样性来实现. 合理设置任务难度是确保认知任务信度的重要因素. 任务过易或过难可能引发天花板或地板效应, 掩盖个体能力差异[52]. 因此, 设计认知任务时, 应确保任务难度与被试群体平均能力水平相匹配[51]. 在能力水平未知的情况下, 设计难度跨度较大的任务或筛选出可能引起极端分数的任务模块, 是增加被试间变异性的有效方法[52,53,74].

提高被试的参与度和投入感, 是增强任务信度的另一个重要策略. 游戏化通过融入游戏设计元素(如角色扮演、叙事背景、互动反馈机制等), 使传统任务呈现游戏特征[55]. Kucina等人[54]的研究表明, 游戏化可以增加任务复杂性, 激发多样化反应, 从而提高信度. 此外, 整合游戏元素或视频刺激可提升参与度和动机, 减轻疲劳效应[53]. 然而, 游戏化并非总是有效. 其对信度的提升在不同研究中表现不一, 且在部分情境中反而可能引入额外的噪音. 例如, 若被试对情节设定不感兴趣, 或对任务流程缺乏投入, 可能导致更大的反应变异和数据不稳定[54,55]. 此外, 近期研究指出[75], 游戏化任务中为增强趣味性所加入的视觉或奖励元素(如音效、动画等)可能会干扰核心心理机制的测量, 从而削弱测量的内部一致性与效度. 个体在游戏经验、策略运用等方面的差异, 也可能导致被试间非系统性变异, 进一步影响测量信度. 因此, 未来在应用任务游戏化的同时, 应注重游戏元素与目标心理机制的深度整合, 并加强对游戏化潜在干扰因素的控制和评估.

此外, 样本的同质性也是影响被试间变异性的重要因素. 若样本来源过于单一, 被试间表现趋于一致, 个体差异难以被有效捕捉. 因此, 研究者可通过扩大样本量[76]以及通过招募社区或在线平台样本(如Amazon Mechanical Turk、Prolific)等方式, 提高被试群体的异质性[51,56]. 不过, 在增加样本多样性的同时, 也必须确保任务在不同群体间具备测量等值性(measurement invariance), 否则群体间数据差异可能反映的并非真实心理结构的差异, 而是测量工具在不同文化或背景下的结构不一致[77]. 例如, 许多心理学研究基于WEIRD (western, educated, industrialized, rich, democratic)群体[57,78,79], 研究者需要采用测量等值性检验(如多组结构方程模型或项目反应理论分析), 检验任务在各群体间的结构、载荷与截距等参数是否具备等值性, 以确保信度评估和变量比较的有效性与可解释性[80].

4.2.2 降低误差分数方差

试次噪音过高会掩盖个体真实差异, 必须采取措施降低测量误差的干扰. 降低误差分数方差的具体策略包括增加正式实验的试次数量、设置充分练习以及严格控制实验环境. 增加试次数量是提高测量分数方差的有效手段. 例如, 在反应时任务中, 增加试次可稳定被试的任务表现并扩大个体差异[64,81]. 帮助更加精准地测量个体的认知能力差异. Liu等人[37]在自我匹配任务中发现, 试次数与分半信度显著正相关. 研究者可借助Spearman–Brown预测公式[82], 结合预实验数据, 估算达到特定信度水平所需的最小试次数. 然而, 试次数与信度之间并非线性关系[83]. 若试次设置过多, 可能引发被试疲劳或任务熟练所致的学习效应, 反而加剧测量误差. 因此, 在试次数设计中需综合考虑任务时长、认知负荷与被试耐受度, 权衡精度与效率. 研究者需评估信度随试次数变化的边际收益, 以确定最优试次阈值, 从而在确保测量稳定性的同时降低额外误差的引入.

练习环节对测量信度的提升也至关重要. 已有研究表明, 在正式实验前设置充分的练习阶段有助于稳定被试表现并降低误差方差. Collie等人[65]指出, 练习有助于被试熟悉任务流程, 从而减少随机波动; McLean等人[64]进一步建议, 设置独立的练习模块, 并将练习数据单独分析或予以剔除, 以避免其影响正式数据.

实验环境的控制对信度评估具有重要意义. 在实验室环境中, 需标准化物理条件(如光照、噪音、座椅高度); 在线测试中, 可以通过注意力检查(如眼动追踪、反应时筛选)识别并排除低质量数据, 增强数据可靠性[84]. 此外, 虚拟现实(VR)技术通过提供沉浸式环境和严格实验控制, 减少外界干扰, 提高生态效度[66]. 值得注意的是, 对于涉及不同设备(如智能手机与桌面端)的测量任务, 需进行测量等值性检验[84].

4.3 改良统计模型方法

鉴于认知任务数据固有的层级结构(试次嵌套于被试)以及传统统计方法在处理这些数据可能存在局限性, 继续改良统计模型方法仍然是提升信度评估的准确性的有效方法, 包含层级模型及改良信度估计方法等策略. 鉴于认知任务数据的层级结构, 分层模型(hierarchical models)被广泛推荐用于区分试次间与个体间变异, 减少试次变异对个体差异估计的干扰[60]. 此外, 分层模型还可能提高结果的可推广性, 增强不同实验设计结果的可比性[60]. 方法上, 贝叶斯分层模型(Bayesian hierarchical models)用于信度估计时, 可以避免经典统计中参数估计的问题[58,59]. 当数据结构更复杂时, 研究者可以考虑使用交叉混合效应模型(crossed random effects model)和混合效应位置尺度模型(mixed-effects location scale model, MELSM)[85,86]. 但分层模型的应用也存在局限, 如贝叶斯模型对先验敏感[87], 且模型参数众多需较大样本量支持, 增加实际操作难度.

除了统计模型的改进, Xu和Stocco[63]创新性地引入非参数信度指标(nonparametric reliability indices)和多变量信度计算方法(multivariate reliability estimation methods), 突破了传统模型对分布和参数结构的依赖. 非参数信度指标如个体区分度(discriminability)衡量个体在不同测量间的变异是否显著大于测量内变异; 指纹法(fingerprinting)则用于评估同一被试多次测量的稳定性. 在多变量信度计算方面, 提出基于距离的ICC (dblICC)和基于信息论的ICC (I2C2). dblICC通过比较个体间与个体内的欧氏距离估算信度, 适用于多维数据; I2C2利用协方差矩阵衡量测量的总体稳定性[76]. Xu等人[12]开发了用于计算这些指标的R包*Rex*, 见网页(https://www.scidb.cn/en/anonymous/QVJqbWFt)中以自我匹配任务的反应时为例[37]的计算代码. 类似地, 为降低对实验设计的依赖, Rouder和Mehrvarz[25,67]提出信号噪声比(signal-to-noise ratio, SNR)作为新型信度评估方法. SNR基于层级模型对个体间变异与试次内噪声的比值进行估计, 提供独立于试次数的信度评估, 较传统系数更直观反映任务的可测性.

4.4 超越经典测量框架

经典测量理论在处理认知任务数据时存在样本依赖性、误差建模能力有限以及难以充分处理多维度信息等局限, 这些局限在追求内部效度而忽视构念效度的背景下尤为突出, 因此有必要超越传统框架, 引入更灵活的现代测量理论视角. 近年来, 一些研究者尝试突破传统经典测量框架, 发展出多种更为灵活的信度评估方法.

现代测量理论的两个重要分支——项目反应理论(item response theory, IRT)和概化理论(generalizability theory, G Theory)为认知任务的信度评估提供了潜在的理论支撑与方法启发. 尽管这两种理论主要应用于问卷和标准化测验, 但其核心理念对于认知任务中信度问题的建模与解释同样具有参考价值. 以项目反应理论为例, 其通过建立潜在能力与个体反应之间的概率关系, 能够揭示不同项目在测量目标特质上的区分度与贡献度[68]. 在认知任务中, 若将每个试次类比为一个“项目”, 理论上可借助项目反应理论框架分析哪些任务条件更能有效捕捉个体差异, 进而优化任务结构, 提高测量效率与稳定性. 同样, 概化理论所强调的多误差源建模视角, 也为理解认知任务中的信度问题提供了更全面的视野. 相比传统信度估计, 概化理论能同时考虑被试间差异、任务内波动、情境变化等多维度误差来源, 这一思想对于认知任务中普遍存在的高误差与低信度问题尤具启发性[69].

5 总结与展望

量化人类认知过程作为科学心理学的核心使命之一, 不仅为实验研究与相关研究范式提供了整合的契机, 也为理解精神障碍的病理机制、开发干预手段开辟了全新路径. 美国国家精神卫生研究所(NIMH)提出的精神疾病研究领域标准(RDoC)框架[88~90]将认知过程确立为关键研究维度, 充分彰显了认知过程在精神健康研究中的战略意义. 然而, 用于测量认知过程的认知任务的信度问题是制约应用的重要瓶颈, 特别是在个体差异研究中, 低信度可能严重影响效应估计的稳定性与结论的可推广性. 因此, 提升认知任务的信度, 已成为推动跨学科认知研究迈向更高精度和更大实用

性的关键一步.

当前, 关于认知任务信度的研究已取得初步成果, 研究者提出了更契合实验设计特点的信度估计方法与新型指标, 并对多种经典任务的信度表现进行了系统评估. 但信度偏低不仅源于测量误差与层级数据结构的复杂性, 更在于任务在衡量个体差异时构念效度的不足. 因此, 未来应从系统建构构念效度、优化任务设计、开发更合适的信度模型, 以及探索超越经典测量理论的新方法等方面着力. 目前有部分策略已获得初步实证支持, 但未来研究仍需构建覆盖多维认知能力的系统性信度评估框架(如[22,25]), 改进并提升现有任务的信度[37,64], 建立适应认知研究的新型评估标准[59,67], 构建大规模基准数据集(如CoRR项目[95]), 并开发面向个体差异研究的新型认知测量范式[54]. 这些探索的推进, 将为揭示人类认知本质规律、推动精神障碍的早期识别与精准干预、促进教育与智能系统的个性化发展, 提供坚实的理论基础和方法支持, 进一步推动心理学和相关学科的交叉融合与创新发展.

参考文献

1 Anderson J R. Cognitive Psychology and Its Implications. London: Macmillan, 2005
2 Deary I J, Penke L, Johnson W. The neuroscience of human intelligence differences. Nat Rev Neurosci, 2010, 11: 201–211
3 Miller G A. The cognitive revolution: a historical perspective. Trends Cogn Sci, 2003, 7: 141–144
4 Kriegeskorte N, Douglas P K. Cognitive computational neuroscience. Nat Neurosci, 2018, 21: 1148–1160
5 Zuo X N, Li H J, Ma H L. Developmental population neuroscience: embracing diversity (in Chinese). Chin Sci Bull, 2024, 69: 3479–3483 [左西年, 李会杰, 马海林. 发展人口神经科学: 拥抱多样性. 科学通报, 2024, 69: 3479–3483]
6 Ou J, Wu Y, Liu J, et al. Computational psychiatry: a new perspective on research and clinical applications in depression (in Chinese). Adv Psychol Sci, 2020, 28: 111–127 [区健新, 吴寅, 刘金婷, 等. 计算精神病学: 抑郁症研究和临床应用的新视角. 心理科学进展, 2020, 28: 111–127]
7 Huys Q J M. Computational Psychiatry. New York: Springer, 2015. 775–783
8 Montague P R, Dolan R J, Friston K J, et al. Computational psychiatry. Trends Cogn Sci, 2012, 16: 72–80
9 Geng S, Liu S, Fu Z, et al. Recommendation as Language Processing (RLP): a Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). RecSys '22: Sixteenth ACM Conference on Recommender Systems. Seattle: ACM, 2022. 299–315
10 Huys Q J M, Browning M, Paulus M P, et al. Advances in the computational understanding of mental illness. Neuropsychopharmacology, 2021, 46: 3–19
11 Lord F M, Novick M R, Birnbaum A. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968
12 Xu T, Kiar G, Cho J W, et al. ReX: an integrative tool for quantifying and optimizing measurement reliability for the study of individual differences. Nat Methods, 2023, 20: 1025–1028
13 Crocker L, Algina J. Introduction to Classical and Modern Test Theory. County of Orange: ERIC, 1986
14 Dang J, King K M, Inzlicht M. Why are self-report and behavioral measures weakly correlated? Trends Cogn Sci, 2020, 24: 267–269
15 Parsons S, Kruijt A W, Fox E. Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. Adv Methods Practs Psychological Sci, 2019, 2: 378–395
16 Yarkoni T, Braver T S. Cognitive Neuroscience Approaches to Individual Differences in Working Memory and Executive Control: Conceptual and Methodological Issues. New York: Springer, 2010. 87–107
17 Vasey M W, Dalgleish T, Silverman W K. Research on information-processing factors in child and adolescent psychopathology: a critical commentary. J Clin Child Adolesc Psychol, 2003, 32: 81–93
18 Hu C, Wang F, Guo J, et al. The replication crisis in psychological research (in Chinese). Adv Psychol Sci, 2016, 24: 1504 [胡传鹏, 王非, 过继成思, 等. 心理学研究中的可重复性问题: 从危机到契机. 心理科学进展, 2016, 24: 1504]
19 Baker M. 1,500 Scientists Lift the Lid on Reproducibility. London: Nature Publishing Group, 2016
20 Schlegelmilch R. Estimating the reproducibility of psychological science. Science, 2015, 349: aac4716
21 Elliott M L, Knodt A R, Ireland D, et al. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol Sci, 2020, 31: 792–806
22 Enkavi A Z, Eisenberg I W, Bissett P G, et al. Large-scale analysis of test–retest reliabilities of self-regulation measures. Proc Natl Acad Sci USA, 2019, 116: 5472–5477
23 Hedge C, Powell G, Sumner P. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav Res, 2018, 50: 1166–1186

24 Huang Y, Luan S, Wu B, et al. Impulsivity is a stable, measurable, and predictive psychological trait. Proc Natl Acad Sci USA, 2024, 121: e2321758121
25 Karvelis P, Paulus M P, Diaconescu A O. Individual differences in computational psychiatry: a review of current challenges. Neurosci BioBehav Rev, 2023, 148: 105137
26 Cronbach L J. The two disciplines of scientific psychology. Am Psychol, 1957, 12: 671–684
27 Vazire S, Schiavone S R, Bottesini J G. Credibility beyond replicability: improving the four validities in psychological science. Curr Dir Psychol Sci, 2022, 31: 162–168
28 Flake J K, Pek J, Hehman E. Construct validation in social and personality research. Soc Psychol Personality Sci, 2017, 8: 370–378
29 Shu Y, Shi Y, Yuan Y. An “operational definition” and a “falsifiability criterion” are not sufficient to lay the foundation for scientific psychology (in Chinese). Acta Psychol Sin, 2019, 51: 1068–1078 [舒跃育, 石莹波, 袁彦. “操作性定义”和“证伪标准”不足以为心理学奠基. 心理学报, 2019, 51: 1068–1078]
30 Sabb F W, Bearden C E, Glahn D C, et al. A collaborative knowledge base for cognitive phenomics. Mol Psychiatry, 2008, 13: 350–360
31 Sui J, He X, Humphreys G W. Perceptual effects of social salience: evidence from self-prioritization effects on perceptual matching. J Exp Psychol-Hum Perception Performance, 2012, 38: 1105–1117
32 Liu Y, Hu C. Behavioral and cognitive neuroscience findings regarding assumptions of the evidence accumulation model (in Chinese). Chin Sci Bull, 2023, 69: 1068–1081 [刘逸康, 胡传鹏. 证据积累模型的行为与认知神经证据. 科学通报, 2023, 69: 1068–1081]
33 Kahveci S, Bathke A C, Blechert J. Reaction-time task reliability is more accurately computed with permutation-based split-half correlations than with Cronbach’s alpha. Psychon Bull Rev, 2024, 32: 652–673
34 Pronk T, Molenaar D, Wiers R W, et al. Methods to split cognitive task data for estimating split-half reliability: a comprehensive review and systematic assessment. Psychon Bull Rev, 2022, 29: 44–54
35 Novick M R, Lewis C. Coefficient alpha and the reliability of composite measurements. Psychometrika, 1967, 32: 1–13
36 Ivanov Y, Theeuwes J, Bogaerts L. Reliability of individual differences in distractor suppression driven by statistical learning. Behav Res, 2023, 56: 2437–2451
37 Liu Z, Hu M, Zheng Y, et al. A multiverse assessment of the reliability of the self-matching task as a measurement of the self-prioritization effect. Behav Res, 2025, 57: 37
38 Zhang Z, Yang L Z, Vékony T, et al. Split-half reliability estimates of an online card sorting task in a community sample of young and elderly adults. Behav Res, 2023, 56: 1039–1051
39 Pronk T. Splithalfr: extensible bootstrapped split-half reliabilities. R package version, 2020, 2: 12
40 Bruton A, Conway J H, Holgate S T. Reliability: what is it, and how is it measured? Physiotherapy, 2000, 86: 94–99
41 Koo T K, Li M Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropractic Med, 2016, 15: 155–163
42 McGraw K O, Wong S P. Forming inferences about some intraclass correlation coefficients. Psychol Methods, 1996, 1: 30–46
43 Psych: Procedures for Psychological, Psychometric, and Personality Research. Version 2.4.12. Evanston, Illinois: Northwestern University, 2024
44 Cicchetti D V, Sparrow S A. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic, 1981, 86: 127–137
45 Kupper L L, Hafner K B. On assessing interrater agreement for multiple attribute responses. Biometrics, 1989, 45: 957–967
46 von Bastian C C, Blais C, Brewer G, et al. Advancing the understanding of individual differences in attentional control: theoretical, methodological, and analytical considerations. https://osf.io/x3b9k
47 Sun S T, Wang N, Wen J H, et al. Adataset of cognitive ontology for neuroimaging studies of self reference (in Chinese). China Sci Data, 2023, 8: 175–189 [孙淑婷, 王楠, 温佳慧, 等. 自我参照的神经成像认知本体论数据集. 中国科学数据, 2023, 8: 175–189]
48 Eisenberg I W, Bissett P G, Zeynep Enkavi A, et al. Uncovering the structure of self-regulation through data-driven ontology discovery. Nat Commun, 2019, 10: 2319
49 Wulff D U, Mata R. Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. Nat Hum Behav, 2025, 9: 944–954
50 Zorowitz S, Niv Y. Improving the reliability of cognitive task measures: a narrative review. Biol Psychiatry-Cogn Neurosci NeuroImag, 2023, 8: 789–797
51 Kyllonen P, Hartman R, Sprenger A, et al. General fluid/inductive reasoning battery for a high-ability population. Behav Res, 2019, 51: 507–522
52 Oswald F L, McAbee S T, Redick T S, et al. The development of a short domain-general measure of working memory capacity. Behav Res, 2015, 47: 1343–1355
53 Allen K, Brändle F, Botvinick M, et al. Using games to understand the mind. Nat Hum Behav, 2024, 8: 1035–1043
54 Kucina T, Wells L, Lewis I, et al. Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. Nat Commun, 2023,

14: 2234
55 Sailer M, Hense J U, Mayr S K, et al. How gamification motivates: an experimental study of the effects of specific game design elements on psychological need satisfaction. Comput Hum Behav, 2017, 69: 371–380
56 Arnon I. Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. Behav Res, 2020, 52: 68–81
57 Henrich J, Heine S J, Norenzayan A. The weirdest people in the world? Behav Brain Sci, 2010, 33: 61–83
58 Pan W K, Wen X J, Jin H Y. Bayesian mixed-effects models: a primer with brms (in Chinese). Psychol Technol Appl, 2023, 11: 577–598 [潘晚坷, 温秀娟, 金海洋. 贝叶斯混合效应模型: 基于brms的应用教程. 心理技术与应用, 2023, 11: 577–598]
59 Haines N, Sullivan-Toole H, Olino T. From classical methods to generative models: tackling the unreliability of neuroscientific measures in mental health research. Biol Psychiatry-Cogn Neurosci NeuroImag, 2023, 8: 822–831
60 Rouder J N, Haaf J M. A psychometrics of individual differences in experimental tasks. Psychon Bull Rev, 2019, 26: 452–467
61 Heaton R K, Akshoomoff N, Tulsky D, et al. Reliability and validity of composite scores from the NIH toolbox cognition battery in adults. J Int Neuropsychol Soc, 2014, 20: 588–598
62 Sullivan-Toole H, Haines N, Dale K, et al. Enhancing the psychometric properties of the iowa gambling task using full generative modeling. Comput Psychiatry, 2022, 6: 189–212
63 Xu Y, Stocco A. Recovering reliable idiographic biological parameters from noisy behavioral data: the case of basal ganglia indices in the probabilistic selection task. Comput Brain Behav, 2021, 4: 318–334
64 McLean B F, Mattiske J K, Balzan R P. Towards a reliable repeated-measures beads task for assessing the jumping to conclusions bias. Psychiatry Res, 2018, 265: 200–207
65 Collie A, Maruff P, Darby D G, et al. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. J Int Neuropsychol Soc, 2003, 9: 419–428
66 Bruder L R, Scharer L, Peters J. Reliability assessment of temporal discounting measures in virtual reality environments. Sci Rep, 2021, 11: 7015
67 Rouder J N, Mehrvarz M. Hierarchical-model insights for planning and interpreting individual-difference studies of cognitive abilities. Curr Dir Psychol Sci, 2024, 33: 128–135
68 Embretson S E, Reise S P. Item Response Theory for Psychologists. London: Psychology Press, 2013
69 Brennan R L. Variability of Statistics in Generalizability Theory. New York: Springer, 2001. 179–213
70 Rappaport B I, Shankman S A, Glazer J E, et al. Psychometrics of drift-diffusion model parameters derived from the Eriksen flanker task: reliability and validity in two independent samples. Cogn Affect Behav Neurosci, 2025, 25: 311–328
71 Hitchcock P F, Fried E I, Frank M J. Computational psychiatry needs time and context. Annu Rev Psychol, 2022, 73: 243–270
72 Eckstein M K, Master S L, Xia L, et al. The interpretation of computational model parameters depends on the context. eLife, 2022, 11: e75474
73 Rouder J N, Kumar A, Haaf J M. Why many studies of individual differences with inhibition tasks may not localize correlations. Psychon Bull Rev, 2023, 30: 2049–2066
74 Feldt L S. The relationship between the distribution of item difficulties and test reliability. Appl Measurement Education, 1993, 6: 37–48
75 Luo W, Luo C, Yan Z, et al. Resting-state fMRI and population neuroscience: progresses and guidelines for reliability research (in Chinese). Chin Sci Bull, 2024, 69: 3547–3559 [罗伟, 罗崇静, 颜志雄, 等. 静息态功能磁共振成像与人口神经科学: 信度研究进展与指南. 科学通报, 2024, 69: 3547–3559]
76 Molenaar D, Feskens R. Relating violations of measurement invariance to group differences in response times. Psychol Methods, 2024, doi: 10.1037/met0000655
77 Liu W, Chen Z, Hu C P. Sample representativeness in psychological and brain science research (in Chinese). Chin Sci Bull, 2024, 69: 3515–3531 [刘伟彪, 陈志毅, 胡传鹏. 心理与脑科学研究中的样本代表性. 科学通报, 2024, 69: 3515–3531]
78 Ghai S, Forscher P S, Chuan-Peng H. Big-team science does not guarantee generalizability. Nat Hum Behav, 2024, 8: 1053–1056
79 Yarkoni T. The generalizability crisis. Behav Brain Sci, 2022, 45: e1
80 Lee H J, Smith D M, Hauenstein C E, et al. Precise individual measures of inhibitory control. Nat Hum Behav, 2025, doi: 10.1038/s41562-025-02198-2
81 Sanders P F, Theunissen T J J M, Baas S M. Minimizing the number of observations: a generalization of the Spearman-Brown formula. Psychometrika, 1989, 54: 587–598
82 Kadlec J, Walsh C R, Sadé U, et al. A measure of reliability convergence to select and optimize cognitive tasks for individual differences research. Commun Psychol, 2024, 2: 64
83 Pronk T, Hirst R J, Wiers R W, et al. Can we measure individual differences in cognitive measures reliably via smartphones? A comparison of the flanker effect across device types and samples. Behav Res, 2023, 55: 1641–1652
84 Brunton-Smith I, Sturgis P, Leckie G. Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects

location–scale model. J R Statistical Soc Ser A-Stat Soc, 2017, 180: 551–568
85 Williams D R, Martin S R, Rast P. Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. Behav Res, 2021, 54: 1272–1290
86 Katahira K, Oba T, Toyama A. Does the reliability of computational models truly improve with hierarchical modeling? Some recommendations and considerations for the assessment of model parameter reliability. Psychon Bull Rev, 2024, 31: 2465–2486
87 Cuthbert B N. Research domain criteria (RDoC): progress and potential. Curr Dir Psychol Sci, 2022, 31: 107–114
88 Cuthbert B N, Insel T R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med, 2013, 11: 126
89 Insel T, Cuthbert B, Garvey M, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry, 2010, 167: 748–751
90 Zuo X N, Anderson J S, Bellec P, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci Data, 2014, 1: 1–3

Summary for “认知任务的测量信度: 进展与前景”

Measurement reliability of cognitive tasks: current trends and future directions

Pengpeng Zhu[1,2†], Zheng Liu[3†], Chunhua Kang[4*] & Chuan-Peng Hu[1,2*]

[1] *School of Psychology, Nanjing Normal University, Nanjing 210097, China*
[2] *Adolescent Education and Intelligence Support Lab of Nanjing Normal University, Laboratory of Philosophy and Social Sciences at Universities in Jiangsu Province, Nanjing 210097, China*
[3] *School of Humanities and Social Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China*
[4] *Zhejiang Philosophy and Social Science Laboratory for The Mental Health and Crisis Intervention of Children and Adolescents, Zhejiang Normal University, Jinhua 321004, China*
† Equally contributed to this work
* Corresponding authors, E-mail: hcp4715@hotmail.com; akang@zjnu.cn

Cognitive tasks are fundamental tools in experimental psychology and cognitive neuroscience, extensively used to probe cognitive mechanisms and assess dysfunctions across diverse domains. Despite their ability to produce robust group-level effects, recent studies have raised concerns about their low reliability in capturing individual differences. The seemly discrepancy between robust group-level effects and poor individual-level reliability, known as the “reliability paradox”, highlights a critical challenge in the application of cognitive tasks for individual-level inference. The paradox is particularly consequential given the increasing use of cognitive tasks in real-life settings such as clinical diagnostics and personalized intervention. However, existing discussions on this issue remain fragmented and lack a comprehensive framework for understanding its causes and identifying viable solutions.

We summarize the issues surrounding the reliability paradox of cognitive tasks and categorize them into two core challenges. The first pertains to the hierarchical data structure intrinsic to cognitive tasks, where data are nested within trials, blocks, and subjects. The second concerns construct validity: most tasks are developed to test the effectiveness of experimental manipulations rather than to measure well-defined cognitive constructs—those typically of primary interest in individual differences research. Relatedly, a weaker form of the construct validity problem is the variability of indicators used to represent individual differences in cognitive performance. A single task may yield many possible indicators, either direct outcomes (e.g., reaction times, accuracy) or derived metrics (e.g., efficiency, sensitivity). These issues are historical and stem from the lack of communication between experimental and correlational approaches in psychology.

The challenge of hierarchical data structure has received increasing attention in recent years, and new reliability metrics tailored to cognitive tasks have been developed. These include split-half reliability and intraclass correlation coefficients (ICCs). Empirical evidence suggests that permutation-based split-half reliability demonstrates superior robustness by effectively accounting for trial-level variability and task-specific noise. For repeated measures designs, ICC (2,1) and ICC (3,1) are recommended, as they provide complementary insights into the generalizability and sample specificity of task performance. We present a practical guide for estimating the reliability of tasks with hierarchical data.

The second challenge concerns the heterogeneity and arbitrariness of indicators selected from task outcomes to assess individual differences. The reliability of different indicators from the same task often varies significantly. We argue that such heterogeneity and arbitrariness arise from a lack of construct validity: the link between an indicator and the underlying cognitive construct is rarely well-defined.

Given the complexity of the reliability issues in cognitive tasks, improving reliability requires multifaceted efforts. First and most importantly, construct validity should be tested and enhanced. For example, researchers may employ multi-task designs and latent modeling approaches to identify underlying constructs. Computational modeling also offers promise for more accurately capturing cognitive processes. Second, as noted in prior literature, optimizing task design can improve reliability. Strategies such as adjusting difficulty levels, increasing trial counts, incorporating gamification elements, and minimizing environmental noise can enhance measurement precision and between-subject variance. Third, new statistical models for estimating task reliability are needed. Reliability metrics that reflect the multilevel structure of task data (e.g., multilevel modeling, signal-to-noise ratio) should be more widely adopted. Finally, we recommend integrating modern psychometric frameworks, including item response theory and generalizability theory, to model error variance across trials, contexts, and individuals with greater granularity.

cognitive tasks, reliability paradox, reliability, individual differences, inter-individual differences

doi: 10.1360/CSB-2025-0551