arXiv:2603.08856v1 [cs.HC] 9 Mar 2026

解析可解释性：最优组合解的以人为中心的标准

Dominik Pegler[1], Frank Jäkel[2], David Steyrl[1], Frank Scharnowski[1], 和 Filip Melinscak[11]维也纳大学心理学院认知、情绪与方法系[2]达姆施塔特工业大学认知科学中心与心理学研究所

作者注

通信 通讯作者：Dominik Pegler。电子邮箱： dominik.pegler@univie.ac.a t

摘要

算法支持系统通常会返回难以理解的最优解。然而，有效的人机协作需要可解释性。当机器生成的解同样最优时，人类必须从中选择一个，但对于究竟是什么使某个解比另一个解更具可解释性，目前仍缺乏精确的说明。为识别具有可解释性的机器解的结构属性，我们提出了一种实验范式：参与者需要在两个同样最优的物品装箱方案中，选择更容易理解的一个。我们发现，偏好能够可靠地反映解的结构的三项可量化属性：与贪心启发式的一致性、简洁的箱内组成，以及有序视觉呈现。在有序呈现和启发式对齐方面观察到最强的关联，组合简洁性也表现出一致的关联。反应时证据并不一致：只有在启发式差异较大时，参与者才主要表现出更快的反应；而总体的基于网络摄像头的注视并未显示出复杂度的可靠影响。这些结果为最优装箱解中的可解释性提供了一种具体的、基于特征的说明，将解的结构与人类偏好联系起来。通过识别可操作的属性——简洁的组成、有序呈现和启发式对齐——我们的研究结果支持对机器解进行可解释性感知优化与呈现，并为量化现实分配与设计任务中最优性和可解释性之间的权衡勾勒出一条路径。

关键词：人机协作、问题求解、可解释性、装箱问题

解析可解释性：以人为中心的最优组合解标准

引言

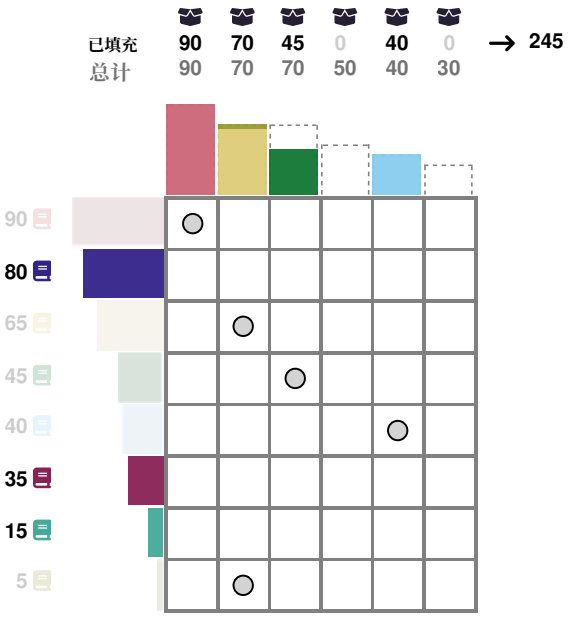
算法优化和机器学习的进步，使自动求解器日益成为人机协作的核心（Akata 等人，2020；Krakowski 等人，2023）。在许多实际应用中，当这些求解器生成计划或分配时，人类可解释性便成为采纳和安全使用的实际前提。许多优化问题都存在多个同样最优、但在结构和呈现方式上差异显著的解。一个尚待解答的研究问题是：当多个最优解的价值相同时，哪些结构属性会让其中一个解比另一个更容易理解？我们以装箱类问题为具体且可控的案例，研究这一普遍的可解释性问题。在这类问题中，多个不同的解可能同样最优，但在人们看来，其易理解程度却可能大不相同。

组合装箱与多重子集和问题（MSSP）

装箱问题——例如经典的装箱问题（Johnson 等，1974）和多重背包（Cacchiani 等人，2022）——要求在硬约束下，将大小各异的物品分配到容量有限的箱中。这类问题是运筹学的基础，在资源分配和物流领域有着重要应用（Gunawan 等人，2021）。例如，医院必须将患者（具有护理需求的物品）分配给数量有限的护士（具有容量限制的箱）（Marzouk 与 Kamoun，2021）。资本预算同样需要在竞争性项目之间分配有限资源（Gurski 等人，2019）。我们研究多重子集和问题（MSSP；Caprara 等人，2000），这是多重背包问题的一个特例：每件物品的利润等于其大小，箱的数量和容量固定，目标是最大化装入总量。多个解可能达到相同的目标值；然而，有些解更易于推理、交流或修改，因此在实践中更有用。图 1 展示了本研究使用的一个

MSSP 实例，并以分配矩阵的形式进行可视化。

图 1多重子集和问题示意图



注：多重子集和问题（MSSP）的一个实例。行表示物品，列表示箱。物品的分配以单元格中的灰点表示。物品尺寸由条块长度和数字标签表示。总体目标分数（装入总量）显示在右上角。

解读最优解

即使呈现得清晰，像 MSSP 这样的组合问题，其最优解在底层结构和依据对人类而言有多容易理解方面仍可能存在显著差异。按照既有用法，我们将这种以人为中心的特质称为可解释性：用户理解机器生成的解（计划或分配）并有效运用它的程度（Doshi-Velez & Kim, 2017）。从心理学角度看，可解释性与知觉、理解和信任相互作用：人们更偏好符合熟悉结构、能够在脑海中模拟或为之辩护的解，即便这意味着放弃不透明但最优的替代方案（Kahneman, 2011; Miller, 2019; Lipton, 2017; A. Tversky & Kahneman, 1974; Dietvorst et al., 2015; Sweller, 1988; Zerilli et al., 2022; Bussone et al., 2015; Lee & See, 2004）。尽管已有大量

可解释人工智能（XAI）研究聚焦于对预测进行解释（Barredo Arrieta 等，2020; Abdul 等，2018; Rudin，2019），但在组合情境中，究竟是什么让某个最优解比另一个更容易理解，我们对此所知甚少（但参见 Ibs 和 Rothkopf，2026; Ibs 等，2024; Ott 和 Jäkel，2023）。关键的是，关于解释可解释性的研究表明，增加解释的复杂度（例如，加入更多术语或新概念）会延长人类核验其一致性所需的时间（Narayanan 等，2018）。

基于复杂度的可解释性代理指标

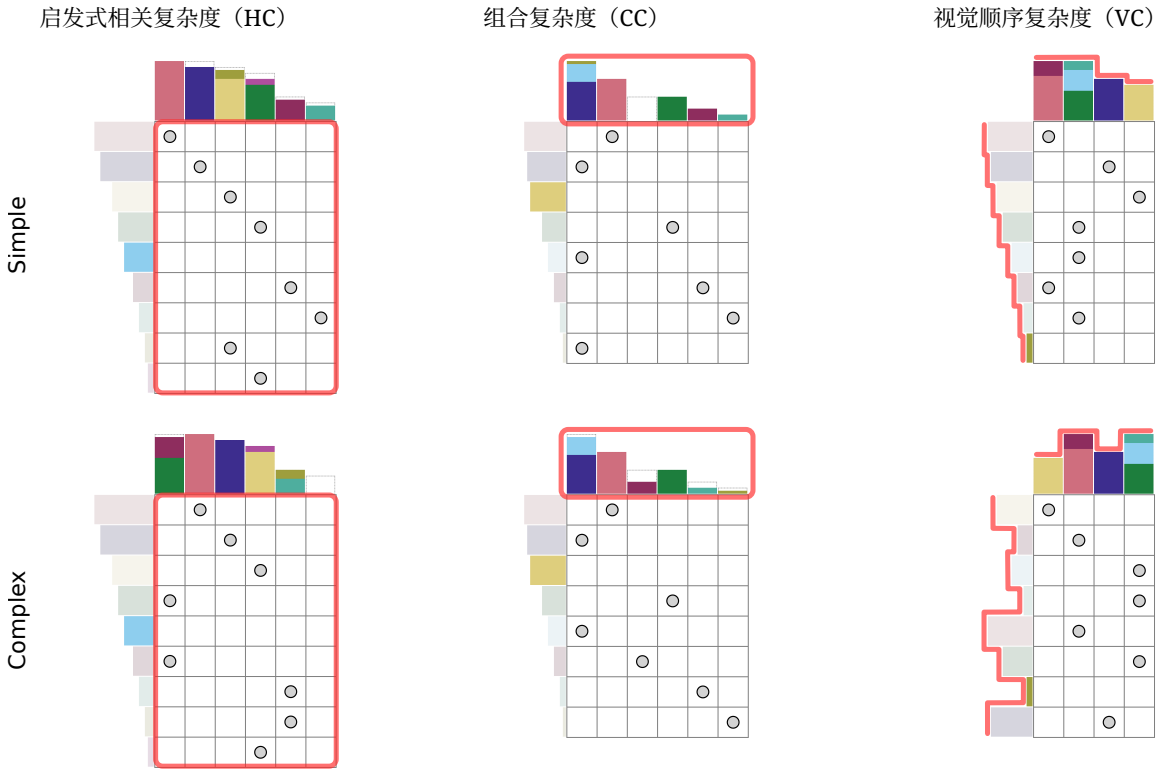
我们聚焦于三种有助于提升可解释性、且符合成熟认知与知觉原理的解级属性。首先，人们通常依靠简单的启发式方法解决问题，更倾向于采用符合熟悉构造规则的结构，并且认为较大的偏离更难以合理解释（Gigerenzer 和 Gaissmaier，2011；A. Tversky 和 Kahneman，1974；Cormen 等人，2009）。其次，组合简洁性能够减轻认知负荷：接近空或接近满、且仅包含少量物品的箱，比装有较多物品的箱或半满的箱更容易编码和比较（Sweller，1988；B. Tversky 等，2002）。第三，知觉组织倾向于有序布局；无论依据简约原则还是经验似然原则，人们都更偏好能够用简短规则（例如“最大者优先”）概括的序列（Feldman，2016；van der Helm，2000；Chater，1996；Helmholtz，1909/1962）。

如图 2 所示，我们通过三个解水平指标来操作化这些属性。此处先概述这些指标，具体定义见方法。启发式相关复杂度 (HC)量化解与贪心装箱启发式之间的偏离程度，用以衡量解在多大程度上遵循直观的构造方式。组合复杂度 (CC)旨在综合考量箱内物品的数量、物品尺寸的均衡程度以及未使用容量，以量化一眼看清箱内物品有多困难；因此，原则上，物品较多且填充程度居中的箱，其复杂度可高于主要由少量物品构成且

几乎空置或几乎装满的箱。视觉顺序复杂度 (VC)用于衡量箱和物品在展示中的无序程度，反映解偏离已排序、具有规则性的呈现方式的程度。作为视觉布局控制变量，我们纳入对角线差异度 (DD)，这是一项协变量，用于捕捉与理想化对角线式分配模式之间纯粹的几何相似性。

图 2

描述解复杂度的三个指标



注。这三个面板描述了我们对三种假设复杂度指标背后的直觉。每个面板聚焦于解的不同方面，如红色标注所示。启发式相关复杂度关注分配以及它们偏离贪心启发式的程度。组合复杂度关注箱以及它们的整洁/有序/填充程度（见定义）。视觉顺序复杂度关注解的元素是否按大小排序。

先前研究与研究空白

可解释规划强调让解与用户的心智模型相契合——通过模型协调、对比式理由或解注释——这突显了

可理解性取决于推导过程和呈现方式（Fox 等人，2017；Chakraborti 等人，2017）。在装箱问题和背包问题研究中，结构规律和贪心启发式已得到充分刻画（Coffman 等，1996；Kellerer 等人，2004）。行为研究表明，人类依赖简单策略，而 performance 取决于实例结构（MacGregor 和 Chu，2011；Dumnić 等人，2019；Murawski 和 Bossaerts，2016；Franco 等人，2021；Franco 等人，2022年；Ibs 等，2024）。即使被要求区分最优程度各异的解，人们也可能难以稳定地识别真正最优的选项，这表明评估复杂组合输出本身就存在困难（Kyritsis 等人，2022年）。

尽管可解释规划提供了证明计划合理性并使其与用户心智模型相契合的方法，且可解释人工智能（XAI）领域在预测性解释方面已取得显著进展（Barredo Arrieta 等，2020；Rudin，2019），但针对组合优化解的可解释性，基于实证和特征的研究仍处于起步阶段（Ibs 等，2024）。与此同时，关于连续优化任务中累积文化演化的研究表明，人们会逐渐偏好并复现符合其归纳偏差的解——即对“好”解应有何种样貌的先验预期或偏好，例如简洁性或对称性；而这些偏差与真正最优解之间的不一致，会系统性地限制集体表现（Thompson & Griffiths，2021）。其他组合领域中关于竞争解的相关研究同样考察了人们的偏好，但并未明确指定解级结构指标（Kyritsis 等，2022）。本研究通过量化三种解级属性——HC、CC 和 VC——如何预测人们在比较同样最优的装箱方案时的选择、反应速度和注意力，来弥补这一空缺。

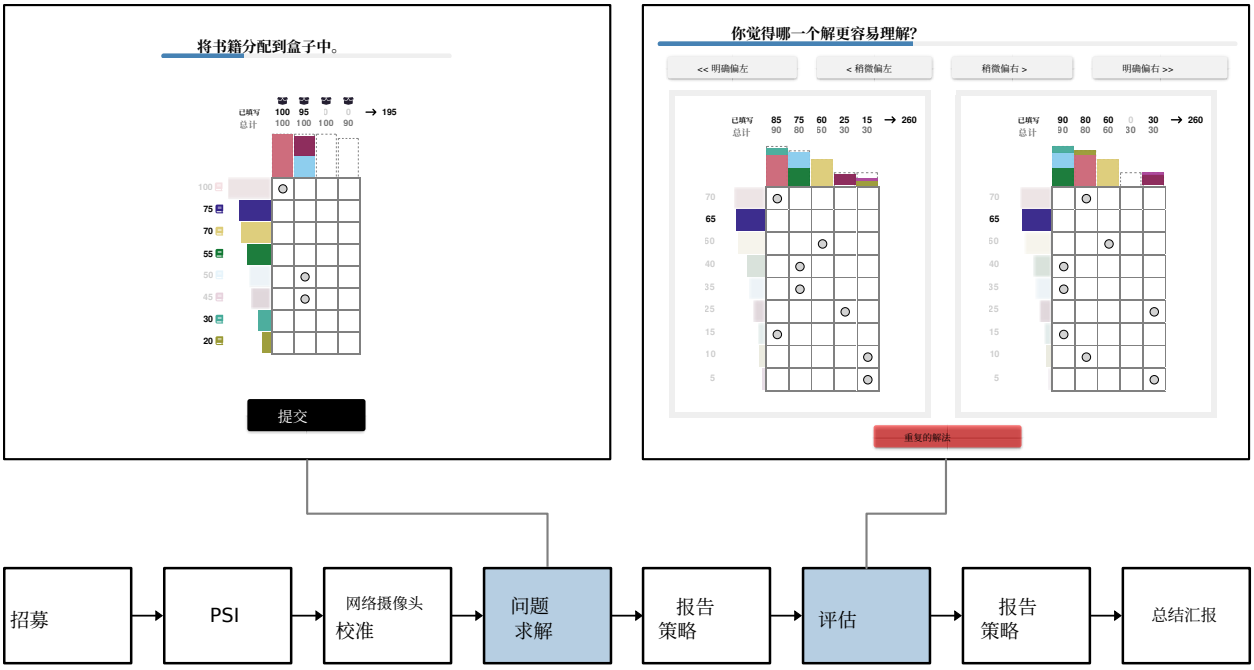
除了描述塑造可解释性的特征之外，我们的实际目标是实现可解释性感知优化。将复杂度指标作为次要准则嵌入——例如，在等值最优解之间进行平局裁决，或在多目标建模中施加软惩罚——将使优化器返回既具有高价值又易于

理解的解（Ehrgott，2005年）。

预注册设计、假设和分析计划

我们使用上文介绍的多重子集和问题研究可解释性（Caprara 等人，2000年；Johnson 等，1974）。每个问题实例都包含若干箱和物品，且经过构造，至少存在两个不同的最优解。参与者首先练习解题，并获得反馈。在主要评估阶段，他们并排查看同一问题实例的两个同样最优的解，并在四级量表（肯定偏左／稍微偏左／稍微偏右／肯定偏右）上回答“哪个解更容易理解？”，从而直接测量其可解释性偏好。图3概述了整个流程。

图3实验流程



注。图示展示了研究流程，包括问卷（PSI = 问题解决量表；Heppner 和 Petersen，1982）、网络摄像头校准、七个带反馈的问题解决试次，以及二十五个评估试次。左侧屏幕显示了一个问题解决试次的示例。在评估试次中（右侧屏幕），参与者判断两个最优解中哪一个更容易理解（“肯定/稍微”偏向左侧或右侧）。

我们检验了三种解级属性是否会影响可解释性偏好：HC、CC 和 VC，并将 DD 作为视觉布局控制变量（完整定义见“方法”部分）。由于参与者评估的是同样最优的解对，我们的分析使用根据这些解级指标得出的解间差异：以带符号的右侧–左侧差值作为选择和注视的预测变量（用于预测哪个选项更受偏好或受到更多查看），以绝对差值分析反应时（用于检验差异较大是否会加快决策）。除偏好外，我们还使用两种过程指标：反应时，用于反映总体处理投入和决策冲突（Luce，1986）；以及基于网络摄像头的注视。基于网络摄像头的眼动追踪可提供汇总的停留时间指标，反映参与者对左侧解与右侧解的相对注意力（Papoutsaki等人，2016；Eckstein 等人，2017；Gollan & Raggam，2025）。这种研究框架使我们能够将可解释的刺激层面结构直接关联到行为偏好、处理速度和注意力。

我们开展了一项探索性研究来完善指标，随后又使用固定指标开展了一项预注册验证性研究。我们假设，在一对同样最优的解中，参与者会更偏好 HC、CC 和 VC 较低的选项；绝对差值越大，决策速度越快；而且复杂度较高的解会吸引相对更多的停留时间。在“结果”部分，我们报告了验证性样本的发现，并将其与探索性样本的发现进行比较。

方法

多重子集和问题（MSSP）

如前所述，我们的研究聚焦于多重子集和问题（MSSP；Caprara 等人，2000年）。该问题是多重背包问题（Cacchiani 等人，2022年）的一种变体，其中每件物品的收益等于其尺寸。给定容量为 w_i 的 m 个箱，以及尺寸为 z_j 的 n 件物品，任务是选择一部分物品，并将每件物品分配到至多一个箱中，同时确保不超过任何箱的容量，并使装入总量最大化，如公式1所示：

$$
\begin{aligned}
\underset{x_{ij}\in\{0,1\}}{\text{maximize}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} z_j x_{ij}, \\
\text{subject to} \quad & \sum_{j=1}^{n} z_j x_{ij} \le w_i, \qquad i = 1, 2, \ldots, m, \\
& \sum_{i=1}^{m} x_{ij} \le 1, \qquad\quad j = 1, 2, \ldots, n.
\end{aligned} \tag{1}
$$

这里， x_{ij} 表示若物品 j 被放入箱 i 则等于 1，否则等于 0。这种箱数固定的最大化目标，不同于经典装箱目标——使箱数最少。图 1 展示了装箱问题变体的一个实例及其实验表示。行对应物品，列对应箱。方块长度和标签表示物品尺寸 z_j，分配矩阵中的填充单元格（圆点）表示分配情况 $x_{ij} = 1$。右上方显示的数字是当前得分，即目标值 $\sum_{i,j} z_j x_{ij}$（见公式1）。

在实验中，我们随机生成了一大组满足若干约束的问题实例。每个问题实例都包含4至6个箱和7至9件物品。此外，所有问题实例均满足以下条件：(1) 没有物品的尺寸大于最大的箱子；(2) 没有箱子的容量小于最小的物品；(3) 物品尺寸总和与箱子容量总和的比值介于0.8至1.0之间；(4) 至少存在两个不同的最优解。选择这些约束是为了构建一组简单但并非毫无挑战的问题。具体而言，将箱子的数量设为4至6个、物品的数量设为7至9件，有助于减少对称性（尤其是考虑到物品尺寸总和与箱子总容量近似呈一一对应关系），并降低最优解恰好对应物品与箱子一一映射的可能性——这种解过于简单，无论是寻找还是评估都缺乏挑战。此外，我们还假设，在资源分配和调度等许多现实应用中，物品数量通常多于箱子数量（Kellerer 等人，2004；Cacchiani 等人，2022年）。关于问题实例及其最优解的生成方式，详见附录A。

实验设计概述

如引言所述，我们采用基于网络的被试内设计，包含两项研究：一项探索性研究，用于提出假设；一项预注册验证性研究，用于检验这些假设。研究遵循《赫尔辛基宣言》，并经维也纳大学伦理委员会批准（IRB 编号：01073）。

参与者提供知情同意后，完成了问题解决量表（PSI；Heppner 和 Petersen，1982），并阅读了详细说明，其中包括一个交互式示例，演示我们的装箱任务（见图 3）。参与者确认自己理解了这些任务（预计耗时约 30 分钟），且网络摄像头眼动追踪完成校准后，实验任务便开始。参与者首先完成七个 问题解决试次，每次试次后都会收到反馈，以帮助他们理解任务。随后，他们在一个自由文本框中报告这一阶段所使用的问题解决策略。接下来，参与者完成 4 个练习评估试次，之后进行 25 个正式评估试次，根据可解释性报告自己对两种解的偏好。然后，他们在另一个自由文本框中报告自己的评估策略。最后，参与者进入总结汇报页面，填写人口统计信息并报告任何与研究相关的问题。

实验流程

参与者

研究通过 Prolific.com（Palan 和 Schitter，2018）招募居住在美国或英国、且符合以下条件的参与者（由 Prolific 根据参与者的自我报告确认）：英语流利；视力正常或矫正后正常；拥有并愿意使用网络摄像头或内置摄像头。参与者的报酬为每小时 9.00 英镑。

在探索性研究（见附录C）中，我们纳入了73名参与者，获得了1,664个观测值（评估试次），并发现复杂度与

可解释性偏好之间存在显著关联，因此验证性研究采用了相同的样本量。我们从 Prolific 招募了87名参与者完成验证性研究，排除后剩余73名（排除率 = 16.1%）。年龄范围为20至78岁（M = 45.00，标准差 = 12.65），样本中男性参与者占60.27%，女性参与者占39.73%。参与者完成实验的中位时长为24.72分钟（第25至第75百分位数 = 19.43–31.65分钟）。探索性研究中的参与者均不得参加本次验证性研究。

在注视分析中，仅纳入具有可用的基于网络摄像头的眼动追踪数据的参与者，因此验证性研究中共有70名参与者和1,600个评估试次（参见注视停留时间）。没有任何有效注视样本的参与者（例如因校准或追踪失败所致）仅参与行为分析。

眼动追踪校准

对于基于网络摄像头的眼动追踪，我们使用了开源 JavaScript库 WebGazer.js（Papoutsaki等人，2016）。在实验试次开始前，参与者按照指示多次注视屏幕上的指定点并点击，以完成眼动追踪校准。随后，参与者会收到有关 WebGazer 提供的校准准确度的反馈；如果校准准确度较差，对话框中便会显示建议重新校准的提示。

实验试次

问题解决试次。 为了熟悉我们的装箱问题变体，参与者必须自行解决七个不同的问题实例（见图3）。这些实例对所有参与者都相同，并按难度递增呈现（所有物品尺寸的总和与所有箱子容量的总和之比）。没有时间限制；每次试次结束后，参与者会被告知其解是否为最优解；如果不是，则会将一个最优解与其自己的解并排展示。

评估试次。 为了回答哪些解比其他解更具可解释性这一问题，在25次评估试次中的每一次，参与者都会看到同一问题的一对最优解。参与者必须通过点击位于该对解上方的四个按钮之一来回答问题“你觉得这两个解中哪一个更容易理解？”，这些按钮的标签分别为：明确偏左、稍微偏左、稍微偏右 和明确偏右 （见图3）。没有时间限制。

25 个评估试次中有两个是注意力检测试次，旨在验证参与者的注意力。在这些试次中，两个解完全相同，参与者需要点击解对下方标有“重复的解法”的第五个按钮（重复解法按钮）。为了确保参与者理解如何在注意力检测试次中作答，评估试次开始前安排了一个练习环节。在该环节中，参与者完成了四个练习试次，其中两个是注意力检测试次。每个试次结束后，参与者都会收到关于其回答是否恰当的反馈。如果练习环节未能正确完成，则必须重新进行。

25 个评估试次中有三个是一致性试次，旨在评估参与者判断的一致性。参与者评估了三组关联解对，一致性判断遵循逻辑排序。例如，如果参与者认为第一个解比第二个解更容易理解，第二个解比第三个解更容易理解，而第一个解又比第三个解更容易理解，这就表明他们对这些解对的评估具有一致性。作出一致性判断的参与者比例决定了我们的模型所能捕捉方差的理论上限，因为它代表了由系统性刺激相关因素驱动的方差。

所有参与者的连贯性试次和注意力检测试次均完全相同，并且在实验中的同一时间点呈现；每位参与者的其余20个试次则从可能解对池中随机抽取。有关试次生成方式的详细说明，请参见附录A。

在验证性研究中，我们为最多200名参与者生成了一个包含5,000个评估试次的试次池（详情见附录A）。附录B（图B2和B3）展示了验证性样本（73名参与者的1,668个试次）中各试次主要预测变量的范围和分布。

问卷

我们使用问题解决量表（PSI；Heppner 和 Petersen，1982）评估参与者自我报告的问题解决技能。该量表由 31 个条目组成，采用六点李克特量表评分。该测量指标可用于内省评估问题解决中元认知和反思方面的个体差异。完成问题解决和评估试次区块后，参与者填写自由文本回答框，描述他们完成任务时使用的策略。最后，事后问卷收集了额外的人口统计数据以及参与者对整体体验的反馈，包括愉悦度、兴趣、说明清晰度和研究时长，均通过李克特量表条目进行评估。

测量指标

为使数据层级明确，我们区分四个嵌套变量层级。参与者层面变量对每个人保持恒定（例如，年龄、专业知识）。问题层面变量在每个问题实例上取一个值（例如，物品数量和箱数量）。解对层级变量针对两个解合在一起计算一次，例如它们的最大值、总和或差值——因此由该试次中的两个解共享。解级变量描述该对中的单个解（例如，左侧解的得分、右侧解的格式）。尽管未作为主要测量指标报告，它们是计算下文所呈现的解对层级变量的输入。下文报告的所有测量指标均标注这些层级名称，以便其在数据结构中的位置明确无误。

因变量

选择（解对水平）。结果变量选择使用四个有序类别记录参与者在评估试次中的反应：明确偏左、稍微偏左、稍微偏右和明确偏右。该变量在所有统计分析中被视为有序因子，并按升序编码：明确偏左 $<$ 稍微偏左<稍微偏右 $<$明确偏右。

反应时间（解对水平）。连续变量反应时（RT）指参与者在评估试次中做出选择所需的时间，以毫秒为单位记录，计时范围从刺激呈现到被试反应。分析中使用反应时的自然对数，以正态化反应时数据通常呈现的正偏态分布。

注视偏向（解对水平）。该解对水平指标根据眼动追踪数据，以左右刺激之间注视样本计数的相对差值进行量化。每个试次中，注视样本会被归入左侧（L）或右侧（R）解。统计分析采用二项广义线性混合模型（广义线性混合模型），并对向量（$R,\ L$）使用 logit 链接函数对这些计数建模。描述性报告中，连续偏向值

$$b = \frac{R - L}{R + L}, \tag{2}$$

的计算结果介于 -1 至 1 之间。没有有效注视样本（$R + L = 0$）的试次将从分析中排除。

复杂度模型

我们使用三种不同的解水平指标来操作化复杂度：启发式相关复杂度（HC）、组合复杂度（CC）和视觉顺序复杂度（VC）（见图 2）。下文将详细说明单个解的各项复杂度度量的推导过程。随后，为了将其作为统计模型中的解对水平预测变量，我们计算每个评估试次中所呈现的一对解之间的差值

：对于选择和注视，计算带符号差值（右侧减左侧）；对于反应时间，计算绝对差值。

启发式相关复杂度（HC，解级）。要计算给定解的 HC，我们首先使用最大箱优先、最大物品优先（LBF-LIF）策略（Coffman 等，1996；Johnson 等，1974）构造一个贪心参考解。具体来说，按容量降序排列箱，并按尺寸降序排列物品。然后依次遍历排序后的箱；对于每个箱，贪心地将当前可用且未分配、并且能够放入该箱的最大物品放入其中，直到无法再放入物品为止。若容量或物品尺寸相同，则保留箱和物品在原始输入中的顺序，以打破平局。接着，我们将给定解和贪心解都表示为二分图（箱/物品作为节点；分配作为边），并计算二者的图编辑距离，其中边插入和删除的代价均为单位代价。HC 即所得距离，数值越大表示与贪心参考解的偏离程度越大。在统计分析中，其带符号的右左差值记为ΔHC，绝对差值记为 |ΔHC|。

对角线差异度（DD，解级；控制协变量）。由于有序问题实例（箱和物品按尺寸降序排列）的启发式解通常在分配矩阵中呈现出一条对角线（从左上到右下；见图 2），我们将其与近似对角线之间的图编辑距离（附录A）纳入分析，作为控制协变量。DD 始终根据所显示的分配矩阵计算：如果箱或物品在视觉上经过置换，则将置换后的显示结果与对角线基准进行比较。因此，DD 反映所观察布局的对角线特征。相比之下，HC 是相对于贪心基准定义的，该基准在进行分配前会在内部按尺寸对箱和物品排序，因此不受视觉置换影响。DD 和 HC 结合起来，可以区分对对角线视觉布局（DD）的偏好与对与启发式对齐的结构（HC）的偏好。在统计分析中，其带符号的右左差值记为 ΔDD，绝对差值记为 |ΔDD|。

组合复杂度（CC，解级）。该指标根据每个箱中物品的组成来评估复杂度。我们将每个箱视为生成模型的结果，并量化该结果在模型下有多么令人意外。在此情境下，惊异度越高，复杂度越高，因为这表明结果偏离了模型所规定的预期模式。反之，惊异度较低意味着简洁性较高，说明该箱与这些偏好的模式高度吻合。一个箱由三个属性描述：（a）物品数量，N；（b）相对物品尺寸向量，C，当箱非空时其总和为 1；以及（c）未使用容量比例，E。假设条件独立性成立，联合密度可分解为

$$p(N,\ C,\ E)\ =\ p(N)\ \cdot\ p(C \mid N)\ \cdot\ p(E \mid N). \tag{3}$$

物品数量。 N 服从从零开始的几何分布，$N \sim \text{Geom}(p)$。该分布为较少的物品数量赋予更高的概率；因此，装有许多物品的箱对意外度得分的贡献更大。

物品尺寸的构成。 对于 $N > 1$，向量 C 服从集中参数为 α 的对称狄利克雷分布，$C \sim \text{Dir}(\alpha)$。对于空箱（$N = 0$）和仅装有一件物品的箱（$N = 1$），此项不存在。当 $\alpha > 1$ 时，模型偏好物品尺寸均匀分配；当 $\alpha < 1$ 时，则偏好一件物品占主导、其余几件尺寸很小的情况。一种可选的校正方法会去除完全均匀分配的基准概率，确保意外度（进而复杂度）反映的是相对于偏好模式的偏离，而非单纯形的大小。

空白空间。 未使用的比例 E 服从一个双成分混合分布，该分布在 0 和 1 附近分配相等的概率质量。因此，几乎装满或几乎空置的箱被视为简单，而半满的箱则被认为更令人意外，因而更复杂。这些成分可以采用三种不同形式之一。例如，在一种变体中，我们使用截断正态混合分布，

$$E \sim \tfrac{1}{2}\,\mathrm{Norm}_{(0,1)}(0,\sigma) + \tfrac{1}{2}\,\mathrm{Norm}_{(0,1)}(1,\sigma), \tag{4}$$

以及与之类似的截断拉普拉斯分布和连续伯努利分布混合形式（Loaiza-Ganem 与 Cunningham，2019）。一个共同的尺度参数 σ 控制概率质量在极值附近集中的程度（下文将讨论这些参数的选择）。对于单个箱，负对数概率

$$L \;= -\left[\ln p(N) \;+\; \ln p(C \mid N) \;+\; \ln p(E \mid N)\right] \tag{5}$$

其以奈特为单位量化意外程度，因而也量化了复杂度。在适当的参数设置下，这一表述使我们能够为由少量物品组成的箱分配相对较低的意外程度：这些箱要么几乎为空，要么几乎装满，且其物品尺寸近似于对称组成；而偏离这些简单模式的情况则可以被分配更高的意外程度。随后，将一个解的组合复杂度定义为在该模型下其各箱意外程度的平均值。在统计分析中，其带符号的右左差值记为 ΔCC，绝对差值记为 $|\Delta\mathrm{CC}|$。

优化后的参数。 如上所述，CC 模型包含若干可调参数，这些参数会影响复杂度评估。空白空间的比例可用不同的分布描述，例如截断正态分布、截断拉普拉斯分布或连续伯努利分布。尺度参数控制这些分布中概率质量的集中程度。此外，参数 p 决定几何分布对包含较多物品的箱子的惩罚力度，而参数 α 则通过狄利克雷分布设定物品尺寸组成的对称性偏好。每次实验（探索性或验证性）开始前，都会执行专门的校准程序以确定这些参数（详情见附录A）。探索性研究中使用的参数信息见附录C。对于验证性分析，此程序得出的参数如下：空白空间采用连续伯努利分布，尺度参数为 $\sigma = 0.426$、$p = 0.043$ 和 $\alpha = 0.984$，并进行狄利克雷校正。

视觉顺序复杂度（VC，解级）。 此指标使用经调整的 Kendall τ（秩相关）版本，评估问题实例可视化表示中物品和箱子的无序程度。令 $w = (w_1、\ldots\ldots、 w_m)$ 表示箱子容量，$z = (z_1、\ldots\ldots、 z_n)$ 表示物品尺寸。对于任意序列 $a = (a_1、\ldots\ldots、 a_k)$ $(k \geq 2)$，定义

$$a_{i,asc} = a_i + i\ \varepsilon, \quad a_{i,desc} = a_i + (k - i + 1)\varepsilon \tag{6}$$

其中 $\varepsilon\ = 10^{-5}$。相应的秩相关为

$$\tau_{asc}(a) = \tau(a_{asc},\ (1,\ 2,\ \ldots,\ k)),\ \tau_{desc}(a)\ = \tau(a_{desc},\ (1,\ 2,\ \ldots,\ k)) \tag{7}$$

无序度量化为

$$d(a) = 1\ - max(|\tau_{asc}(a)|, |\tau_{desc}(a)|). \tag{8}$$

最后，解的视觉顺序复杂度（VC）为

$$VC\ = \frac{m\ d(w)\ +\ n\ d(z)}{m\ +\ n}, \tag{9}$$

其中 $m = |w|$ 和 $n = |z|$。

这种调整将尺寸相同的相邻箱或物品视为已经排好序，具体做法是加入一个微小偏移以打破并列（公式6）。随后，将Kendall的 τ 分别对照升序参照和降序参照进行计算，并保留绝对值较大的相关系数（公式7）；该序列的紊乱按公式8定义。将此过程应用于箱序列和物品序列，并按各自数量对两个紊乱评分加权，即可得到视觉顺序复杂度（公式9）。该表达式量化了某一问题实例的给定可视化表示相对于理想有序状态的总体紊乱程度。在统计分析中，其带符号的右左差值记为 Δ视觉顺序复杂度，其绝对差值记为 |Δ视觉顺序复杂度|。

进一步的测量指标

最大无序度（MD，解对水平）。 这一解对水平指标由两个解的视觉顺序复杂度推导而来。它并非取右侧解与左侧解之间的差值，而是定义为 $max(VC_L, VC_R)$。我们特别关注这一调节变量如何影响启发式相关复杂度、组合复杂度和对角线差异。其理由是，紊乱可能会妨碍比较。要使两个实例之间的比较受到妨碍，只需其中一个实例具有使比较变得困难的特征即可。尽管MD与两个解之间的视觉顺序复杂度差值捕捉的方面略有不同，但它们源自相同的数据。因此，在后续模型分析中，我们没有考虑MD与 Δ视觉顺序复杂度以及 |Δ视觉顺序复杂度| 的交互作用，因为我们预计无法从中得出任何有意义的结论。

问题难度（PD，问题层级）。这一问题层级指标以物品尺寸总和与所有箱子容量总和的比值进行操作化。在抽样的问题中，该连续指标的取值范围为 0.8 至 1.0，数值越高表示特定问题实例的难度越大（另见附录A中的问题和解实例的生成）。该比值通过反映待填充空间的相对紧张程度，捕捉了装箱任务本身的挑战。比值越高，意味着物品总体所占空间越接近可用容量，因此对最优装箱策略的需求也越高；这一细微之处与有关资源约束和认知负荷的理论观点相符（Sweller，1988）。

启发式最优性（HO，问题层面）。该指标用于评估针对给定问题实例的启发式解的质量。它是启发式评分与最优评分的比值，取值范围为 0.0 至 1.0。值为 1.0 表示启发式方法得到了最优解，较低的值则表明该启发式方法对于该问题实例效果不佳。

自我报告的问题解决技能（PSI，参与者层面）。该指标通过计算参与者在 PSI 问卷所有物品上的得分总和得出。

问题解决效率（PSE，参与者层面）。该指标评估参与者的难度加权问题解决效率（PSE），综合考量其解的最优性及七个问题解决试次中的反应时间（RT），其中难度较高的试次对最终得分的贡献按比例更大。对于试次 i，未加权效率定义为

$$E_{i,j} = \frac{S_{i,j}}{O_i} \div RT_{i,j}, \tag{10}$$

其中，$S_{i,\ j}$ 是参与者 j获得的得分，O_i 是试次 i中该问题实例的最优评分，而 $RT_{i,\ j}$ 是反应时间（秒）。试次特定难度权重（w_i）根据群体效率得出。令 n 表示问题解决试次的数量[−]（此处为 $n = 7$），令 E_i 表示参与者在试次 i上的 E 均值。随后将难度权重定义为

$$w_i = \frac{1}{n-1}\left(1 - \frac{\bar{E}_i}{\sum_{k=1}^{n} \bar{E}_k}\right), \tag{11}$$

因此，均值效率较低（难度较高）的试次会获得较大的权重。参与者的总体 PSE（η_j）为加权总和

$$\eta_j \ = \sum_{i=1}^{n} w_i \ E_{i,j}. \tag{12}$$

数值越高，表明参与者越能找到更接近最优解且耗时更短的解，尤其是在最具挑战性的问题中。

数据分析

数据排除与预处理

为确保本研究的数据完整性，我们采用了若干排除标准。如果参与者在两次注意力检测试次中均未点击重复解按钮，则将其排除。如果参与者在至少两次非捕捉试次中使用了该按钮，也将其排除。此外，如果参与者点击了重复解按钮，则将相应的单个试次从分析中排除。未完成全部试次的参与者，其数据也将被舍弃。对于注视分析，任何没有刺激区域注视（$R + L = 0$）的试次也将被排除。

为进行统计分析，我们将所有预测变量置于可比较的尺度上。对于三个复杂度指标和对角线差异度，我们首先计算每个试次的原始右–左差值。随后，将这些原始差值除以所有试次中差值的标准差。我们没有减去差值的均值，因此数值 0 仍表示“两个解之间没有差异”。带符号的标准化差异用作选择分析和注视分析的预测变量，绝对标准化差异则用于反应时分析。注视结果变量未进行标准化。绘图和报告中，注视以偏差表示，$b = (R - L)/(R + L)$，范围为 -1 至 1。

线性混合效应模型

对于每项分析，我们在 R v4.3.2 中拟合了混合效应模型（R 核心团队，2023）。预测变量均作为固定效应纳入。所有模型均使用参与者的随机截距；在收敛条件允许的情况下，还为所纳入的复杂度主效应（以及存在时的 DD）设置随机斜率；交互作用不设置随机斜率。附录A详细说明了在候选模型中保持一致的随机效应程序，以及基于赤池信息准则（AIC）的选择流程。该流程比较了一组由理论依据驱动的候选模型。如果最佳拟合模型比所有其他模型至少低 2 个 AIC 单位（$\Delta\text{AIC} > 2$），则选定该模型。当两个或

更多模型与最小值相差不超过 2 个 AIC 单位（$\Delta\text{AIC} \leq 2$）时，则认为它们得到同等支持（Burnham 和 Anderson，2004），并从中选择最简单的模型（参数最少）。有序结果使用 `clmm` 来自 *ordinal* 包的模型进行分析（Christensen，2023；`thresholds = "symmetric"`、`link = logit`，拉普拉斯近似），连续结果使用 `lmer` 来自 lme4 的模型进行分析（Bates 等人，2015；`REML =FALSE`），二项计数使用 `glmer` 来自 lme4 的模型进行分析（`family = binomial`）。对于这三类模型，我们均使用 *performance* 包（Lüdecke 等人，2021）报告 Nakagawa 的边际和条件 R^2。

三项分析均使用 HC、CC、VC 和 DD 中的右–左差值作为核心预测变量，并采用同一组潜在调节变量或协变量：PD、PSI、PSE、HO 和 MD。在所有模型中，MD 从未与 VC 同时纳入，因为二者均源自同一组基础紊乱评分。

三项分析在结果变量、预测变量形式和交互作用处理策略方面有所不同。对于选择，根据带符号的标准化右左差异预测有序结果（四个水平），每个复杂度预测变量与每个调节变量之间均设置二阶交互作用（主效应之间以及与 DD 之间不设置交互作用）。对于 反应时间，使用解之间的绝对差值对连续型结果（反应时对数）进行建模，将协变量仅作为其他协变量纳入，不设置交互作用。对于 注视偏向，基于左右解上的注视样本计数拟合二项广义线性混合模型，根据两侧注视样本总数对注视右侧解的概率进行建模；固定效应为带符号的右侧–左侧差值，并采用与选择模型相同的调节变量集合和交互作用处理策略。

一致性上限估计

我们检查了三个一致性试次（见评估试次），并确认每位参与者的一组成对评分是否具有传递性。符合这一标准的参与者比例（p_{coh}）构成了可归因于刺激属性的方差的经验上限。因此，我们将 p_{coh} 与

GLMM 的边际 Nakagawa R^2，该模型根据解复杂度预测有序选择，因为边际 R^2 仅分离出固定效应（复杂度）所解释的方差，而条件 R^2 还会包括被试随机因素带来的方差，因此会超过理论上可解释的范围（Nakagawa & Schielzeth，2013）。此分析使用 R 软件包$ordinal$（Christensen，2023）和$performance$（Lüdecke 等人，2021）完成。

预注册

在收集数据之前，我们已在 OSF 上为验证性研究预注册了假设、主要结果、预测变量、样本量、排除标准和分析计划（https://doi.org/10.17605/OSF.IO/D2AQ7）。探索性研究早于此次注册，并用于完善指标和刺激材料。与预注册计划的任何偏离均列于下文。

偏离预注册之处

我们采用了两项偏离，并在 OSF 记录中均予以注明。首先，在反应时（RT）分析中，我们加入了自我报告的问题解决技能（PSI；经 z 标准化；（Heppner & Petersen，1982））作为潜在协变量。这纠正了一处疏漏：PSI（自我报告）和 PSE（行为表现；预注册的潜在协变量）反映的是互补的构念，两者都可能影响 RT。这项变更仅影响 RT 模型；选择和注视分析仍按预注册方案进行。在基于赤池信息准则的模型选择中，最终的 RT 模型未纳入 PSI，而将 PSI 作为候选变量纳入后，最终也没有改变显著效应的模式或结论。其次，我们修改了核心预测变量（HC、CC、VC、DD）的标准化程序。根据预注册方案，我们原计划先对合并后的左侧和右侧数值进行 z 标准化，然后计算这些 z 分数之间的差值（右侧–左侧）。在最终分析中，我们改为先计算原始的右–左差值，再将这些差值分数除以其经验标准差进行标准化，不减去均值，以便让 0 保持为一个可解释的参照点（选项之间无差异）。这项变更

仅重新缩放了预测变量（因此也重新缩放了回归系数），没有影响标准误、检验统计量或 p值；显著效应的模式和结论均保持不变。

结果

对较简单解的偏好

我们采用基于赤池信息准则（AIC）的模型选择（表1），使用序数混合效应模型检验启发式相关复杂度（HC）、组合复杂度（CC）和视觉顺序复杂度（VC）的标准化右左差异，以及对角线差异度（DD）。与假设一致，参与者偏好较简单的选项：三种复杂度差异（HC、CC、VC）的系数均为负，而DD无法可靠地预测选择。差异每增加一个标准差，选择较复杂解的优势便分别降低27%（HC；OR $= 0.73$，95% CI$[0.64, 0.83]$）、21%（CC；OR $= 0.79$，95% CI $[0.70, 0.90]$）和31%（VC；OR $= 0.69$，95% CI$[0.62, 0.77]$）。在来自73位参与者的1,668条观测值中，模型拟合度为0.083（边际R^2）和0.201（条件 R^2）；边际 R^2 远低于0.877的经验一致性上限（即反应具有传递性的参与者比例；参见方法部分的一致性上限估计；64位参与者的判断具有传递性，9位参与者的判断具有非传递性），这表明相当一部分选择方差可能归因于系统性因素，而非单纯的决策噪声。图4显示，随着差异增大，预测概率向较简单的选项偏移；图5则展示了具有代表性的刺激对。

在所有评估试次中，选择比例为：明确偏左18.4%，稍微偏左38.0%，稍微偏右29.5%，明确偏右14.1%。核心预测变量之间的两两相关程度适中（最大 $|r| = 0.38$），表明共线性有限（附录B）。

这些验证性研究发现总体上与探索性分析一致。不过，一个值得注意的差异是，DD 的主效应在验证性样本中未达到统计显著。跨研究汇总图见附录D。

表 1预测选择的序数混合效应模型

Term	估计值	SE	z	p	置信区间下限	置信区间上限
固定效应						
中心阈值	0.136	0.061	2.239	0.025	0.017	0.255
阈值间距	1.898	0.054	35.063	$<$ 0.001	1.792	2.004
ΔHC	-0.314	0.064	-4.919	$<$ 0.001	-0.439	-0.189
ΔCC	-0.234	0.066	-3.542	$<$ 0.001	-0.363	-0.104
ΔVC	-0.371	0.057	-6.541	$<$ 0.001	-0.482	-0.260
ΔDD	-0.031	0.061	-0.516	0.606	-0.150	0.087
随机效应						
标准差 (截距 | 受试者)	0.311					
标准差 (Δ启发式相关复杂度 | 受试者)	0.287					
标准差 (Δ组合复杂度 | 受试者)	0.360					
标准差（Δ视觉顺序复杂度 | 受试者）	0.250					
标准差（Δ对角线差异 | 受试者）	0.243					

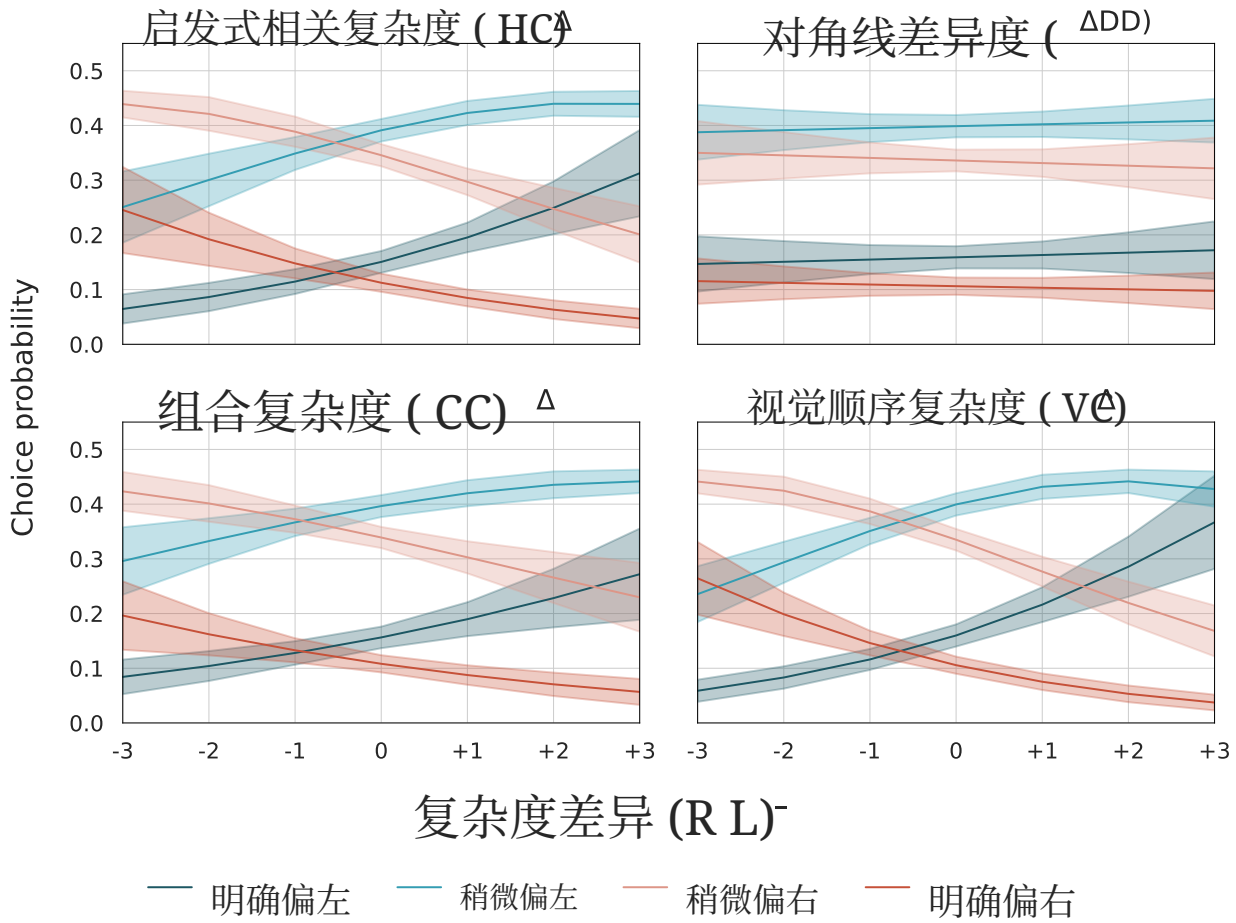
注。 选择 $\sim$ Δ启发式相关复杂度 $+$ Δ组合复杂度 $+$ Δ视觉顺序复杂度 $+$ Δ对角线差异 $+$ （1 $+$ Δ启发式相关复杂度 $+$ Δ组合复杂度 $+$ Δ视觉顺序复杂度 $+$ Δ对角线差异| 受试者）。 N（obs） $= 1668$， N（$subj$） $= 73$，对数似然 $=$-2101.4。预测变量每变化 1 个标准差的优势比（OR）为 OR $=$exp(β)；相应的优势百分比变化为 100 $\cdot$(exp(β) $- 1$)。

反应时间：启发式差异越大，反应越快

我们采用线性混合效应模型，根据 HC、CC、VC 和 DD 的解之间的绝对差值（基于 AIC 的选择），分析了对数反应时（ RT ）。此分析检验了决策速度假设，即更大的差距有助于做出选择。正如假设所预期，更大的 |ΔHC| 可预测更快的反应， |ΔHC|每增加 1 个标准差，RT 平均缩短 4%。相比之下， |ΔCC|、 |ΔVC|和 |ΔDD|均不能显著预测反应；而更高的问题解决效率（PSE）则可以（表 2）。模型拟合度一般

图4

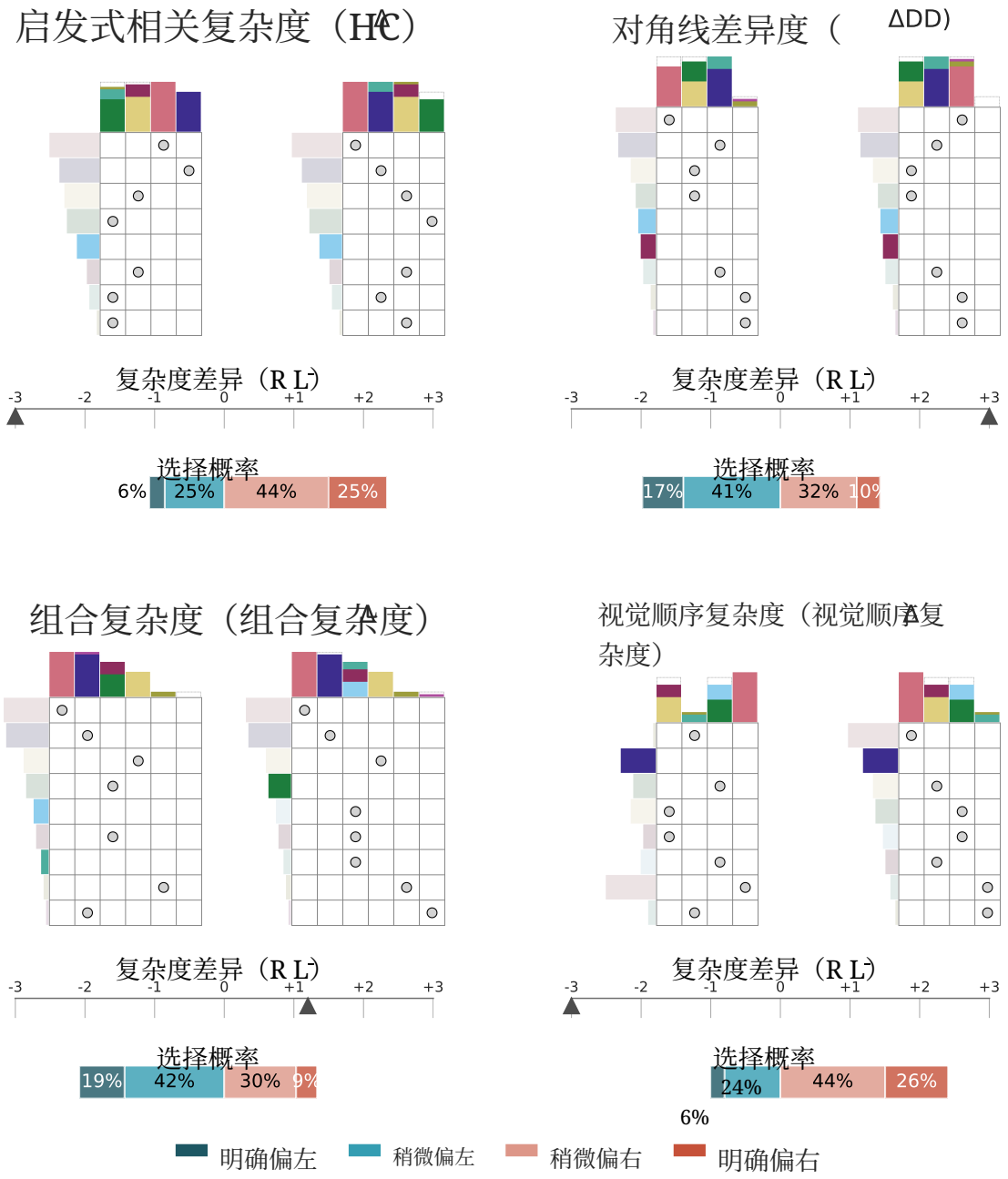
复杂度差异的预测选择概率



注。面板对应三个复杂度指标（HC = 启发式相关复杂度、CC =组合复杂度、VC = 视觉顺序复杂度）以及协变量对角线差异度（DD）。复杂度差异已进行标准化（0 = 无差异； 1 = 一个标准差）。彩色线条表示四种行为反应（“明确偏左”“稍微偏左”“稍微偏右”“明确偏右”）的模型预测概率；阴影带表示 95% 置信区间。

图5

带有模型预测的选择概率的示例配对



注。实验中使用的刺激对示例，涵盖三种复杂度指标（HC、CC、VC）和协变量对角线差异度（DD）。每个面板显示左右两侧的解及其复杂度差值（R–L；负值表示左侧更复杂），并以三角形标记表示。复杂度差异经过标准化（0 = 无差异； 1 = 一个标准差）。下方的水平条显示模型预测的选择概率（%），即明确／稍微倾向选择左侧或右侧的概率（对应图4）。

（边际 $R^2 = 0.054$；条件 $R^2 = 0.649$），涵盖来自73名参与者的1,668个试次。相应原始反应时的中位数为7532毫秒（第25–第75百分位数 = 4708–12954毫秒）。

这些结果与我们的探索性分析形成对比：在该分析中，全部三项复杂度指标——启发式相关复杂度、组合复杂度和视觉顺序复杂度——的绝对差值越大，反应越快。跨研究汇总图见附录D。

表 2预测反应时的线性混合效应模型

Term	估计值	SE	t	p	置信区间下限	置信区间上限
固定效应						
截距	9.010	0.068	131.771	< 0.001	8.874	9.146
|Δ启发式相关复杂度|	-0.042	0.019	-2.214	0.027	-0.079	-0.005
|Δ组合复杂度|	0.016	0.016	1.013	0.311	-0.015	0.048
|Δ视觉顺序复杂度|	-0.004	0.013	-0.325	0.745	-0.030	0.022
|Δ对角线差异|	-0.029	0.016	-1.780	0.075	-0.062	0.003
PSE	-0.167	0.068	-2.460	0.016	-0.303	-0.032
随机效应						
标准差（截距）	0.535					
标准差（残差）	0.411					

注：反应时 $\sim$ |Δ启发式相关复杂度| + |Δ组合复杂度| + |Δ视觉顺序复杂度| + |Δ对角线差异| + PSE + (1 |受试者)。结果变量：对数反应时（反应时对数）。 $N(obs) = 1668$, $N(subj) = 73$, 对数似然 =-1018.1。

没有证据表明复杂度效应会影响注视停留时间

我们对右侧（R）和左侧（L）解的注视样本数量进行二项广义线性混合效应模型（GLMM）分析，以 logit 链接对侧向注视时长建模；等价地，结果变量为 $p = R/(R + L)$。此分析检验复杂度的带符号差值是否能够预测注视停留不对称性。基于AIC的比较保留了仅含截距的设定，表明复杂度对注视

偏向没有可靠影响。截距显著为负（b =-0.400, $p = < 0.001$，95%置信区间（CI） [-0.612, -0.187]），与整体上轻微的左侧注视倾向一致。拟合指标较低（边际 $R^2 = 0.000$；条件 $R^2 = 0.182$）。注视偏向的均值为 -0.062（标准差 $= 0.467$），且无可用注视数据的试次占4.1%。

这些结果与探索性分析一致，后者同样保留了仅截距模型。

讨论

在本文中，我们探讨了装箱方案的哪些属性会使其更易于理解。我们将同一问题的两个最优解并列展示，并收集了参与者的等级偏好。参与者的选择始终倾向于在三项预先定义的指标上复杂度较低的解——组合复杂度（CC）、视觉顺序复杂度（VC）和启发式相关复杂度（HC）。反应时表明，当启发式相关复杂度的差异较大时，决策速度会选择性加快；而汇总的基于网络摄像头的注视数据并未显示出由复杂度驱动的注视时长不对称。综合来看，这些发现支持一种基于特征的最优装箱解可解释性观点，并表明了使机器生成的方案符合人类偏好的实用方法。

可解释的结构：与人类启发式和知觉组织相契合

我们的结果支持主要假设：三种复杂度差异都能预测选择，表明人们稳定地偏好较简单的解。这些相互印证的效应符合一种简明的认知解释。首先，视觉顺序有助于知觉系统生成简短的规则式描述（例如“从最大到最小”），这与日常知觉中的简洁性／似然性原则一致（Feldman，2016；van der Helm，2000；Chater，1996；Helmholtz，1909/1962）。其次，启发式对齐使人们能够立即合理解释某个解是如何构建的，从而减轻解释负担（Gigerenzer 和 Gaissmaier，2011）。第三，组合简洁性降低了编码需求：极端的

箱体组成（接近空或接近满）提供了可一眼捕捉的概括性线索，可能减少进一步知觉加工的需要（Sweller，1988）。

值得注意的是，HC 的稳健效应表明，参与者在评估已完成的解时，可能也会运用熟悉的启发式方法，而不只是生成解时如此。这一观察将启发式方法文献（该领域主要关注解的构建；Gigerenzer 和 Gaissmaier，2011；Cormen 等，2009）拓展至对预计算解的评估。这也与使用欧几里得旅行商问题解的辨别范式研究结果相呼应：在这些研究中，简单的几何性质会引导人们判断哪条巡回路线更优（Kyritsis 等人，2022）。我们的结果表明，与贪心装箱启发式相一致，会系统性地影响人们对同样最优的解的可解释性偏好。将评估判断视为启发式方法的运用，有助于解释为何与我们的参照贪心启发式更一致（HC 较低）的解更容易理解。

更大的启发式差异（|Δ启发式相关复杂度|）与更快的评估相关，这与我们的第二项假设以及熟悉的构建方式能够减少决策冲突这一观点一致（Gigerenzer 和 Gaissmaier，2011；Sweller，1988）。相比之下，箱体组成与顺序的差异（|Δ组合复杂度|和 |Δ视觉顺序复杂度|）并未可靠地缩短决策时间，这表明在我们这种低压力情境下，这些特征虽会引导偏好，却不一定会缩短整体思考时间。尽管探索性样本显示出更广泛的反应时缩短，但此处 HC 特有的模式可能取决于刺激分布和队列差异，因此 HC 是唯一稳健的速度效应（Luce，1986）。

我们假设复杂度差异会表现为注意不对称性；然而，在基于网络摄像头的追踪下，汇总的侧向注视时长并未随复杂度可靠变化，结果提示存在轻微的左侧注视倾向。在同等最优备选方案的配对呈现中，短暂或微小的不对称性可能被试次间变异所淹没。

局限性

我们的研究存在若干局限。首先，我们对可解释性与加工的测量本身受到限制。参与者对偏好诱发提示（“你觉得这两种解中哪一种更容易理解？”）可能有不同理解。选择可能受到视觉吸引力或与个人偏见契合等因素的影响，从而可能将“易于理解”与对某些视觉特征的单纯“喜好”混为一谈（Feldman，2016；van der Helm，2000；Chater，1996；Helmholtz，1909/1962）。然而，启发式对齐（启发式相关复杂度）的一致影响表明，参与者在一定程度上关注了解的结构，而不仅是表面视觉线索。此外，我们使用基于网络摄像头的眼动追踪进行注视测量，空间精度有限，限制了扫描路径等细粒度分析，并可能降低对微妙的、由复杂度驱动的注意力动态的敏感性（Papoutsaki等人，2016）。

其次，我们的实验设置要求参与者在没有时间压力的情况下评判已完全计算好的最优解，这在生态效度上存在权衡。尽管这种受控环境便于进行清晰比较，但它偏离了现实世界中的资源分配与设计任务；这类任务往往涉及部分解、动态约束、风险和截止日期（Lee & See, 2004; Dietvorst et al., 2015）。

第三，尽管我们设计抽样和校准程序是为了系统地变化复杂度，但我们不能排除其他未测量的结构属性与我们的指标共变、并促成所观察到的偏好的可能性。我们的指数捕捉了解的结构中具有理论依据的方面，但它们仍然只是代理指标，可能仅与参与者所敏感的更深层潜在规律不完全相关。这意味着，我们的结果应被解释为启发式相关复杂度、组合复杂度和视觉顺序复杂度是可解释性的信息性标志的证据，而不是证明它们穷尽了相关结构因素的空间。

最后，我们研究结果的普适性受到刺激材料范围的限制：

我们考察了规模相对较小的问题实例（4–6 个箱、7–9 件物品），并基于单一贪心策略（最大箱优先、最大物品优先）定义了启发式相关复杂度。这些结果能否适用于规模更大、复杂度更高的问题，或其他符合人类直觉的启发式方法（Coffman 等，1996；Johnson 等，1974；Kellerer 等人，2004），仍有待确定。

未来方向

未来工作应优先提升测量与生态效度。除主观偏好外，还可开发基于 performance 的评估。例如，在过程层面范式中，参与者完成部分已完成的解，可得到基于任务的解的可用性指标，如准确率和时间。为收集更丰富的主观数据，未来工作可开发或改编专门的问卷，用于感知可解释性、认知负荷和满意度（Brooks et al., 2012; Doshi-Velez & Kim, 2017; Narayanan 等，2018; Afsar et al., 2023），同时从当前数据集的自由文本评估报告分析中获得更深入的见解。与此同时，整合实验室眼动追踪和瞳孔测量，将为与 HC、CC 和 VC 相关的早期注意力分配和认知负荷动态提供更丰富的见解，从而直接应对基于网络摄像头的注视测量固有的局限（Eckstein 等人，2017; Gollan & Raggam, 2025）。为更好地反映现实世界的资源分配与设计任务，在这些实验范式中嵌入时间压力和动态约束，对于提升生态效度至关重要（Lee & See, 2004; Dietvorst et al., 2015）。

推广我们的研究发现是至关重要的下一步。这需要在更广泛的装箱问题和背包问题变体中验证我们的指标，包括规模更大、复杂度更高的问题实例（Kellerer 等人，2004；Cacchiani 等人，2022年；Gurski 等人，2019）。此外，未来研究应探索计算 HC 时除最大箱优先、最大物品优先策略以外，其他符合人类直觉的启发式方法（Coffman 等，1996；Johnson 等，1974；Kellerer 等人，2004）。直接考察呈现策略也

至关重要；这包括将逐步推导（例如，重放解序列或交互式展示）与静态的最终解进行比较。此类研究可以检验展示解序列是否能提升理解，尤其是对于与启发式对齐的解（Fox 等人，2017；Chakraborti 等人，2017；B. Tversky 等，2002）。根据用户的个体偏好，对解的生成和呈现进行个性化，也代表着人机协作的一个有前景的方向（Miller，2019；Zerilli 等人，2022年）。

一项重要的理论与实践挑战，是量化可解释性与最优性之间的权衡。这可以通过在多目标优化公式中将可解释性项纳入次要目标来实现（Ehrgott，2005年）。此类研究有助于确定人们何时更偏好简单但客观上较差的解，并描绘出可解释性可能胜过严格最优性的决策区域。

最终，更长远的目标是开发一种统一的认知模型，将启发式相关复杂度、组合复杂度和视觉顺序复杂度整合为一种概括性的可解释性表征，以解释选择、反应时和注意力。通过样本外预测和生理过程测量指标（如注视、瞳孔测量）验证这一模型，将为理解复杂决策环境中的人类可解释性提供一个全面的框架（Luce，1986；Eckstein 等人，2017；Franco 等人，2021；Franco 等人，2022年）。

结论

在本文研究的组合装箱范式以及可能相关的优化问题中，我们的结果表明，人们对机器生成的可解释解的偏好受三种可量化结构属性的影响：视觉顺序、与贪心启发式的一致性，以及组合简洁性。这些发现为兼顾可解释性的解呈现与优化提供了可行的设计原则。在呈现方面，可以通过对箱和物品进行排序来降低视觉顺序复杂度，从而减少感知混乱。在优化方面，可以将可解释性视为次级准则，例如在质量相当的候选解中优先选择 CC 和 HC 较低的解、在打破平局时倾向于 CC/HC 较低的解、在多目标建模中加入较小的复杂度惩罚项

（Ehrgott，2005年），或筛选一份最优解候选清单，并呈现其中最具可解释性的解。更广泛地说，除传统的性能标准外，将可解释性纳入明确的优化目标，或有助于提高透明度、加快人们作出恰当反应、增强信任，并支持人们在解决问题的人机交互中果断地进行控制。

数据可用性声明

本研究所使用的代码、材料和数据均已公开存放于开放科学框架（OSF）存储库中。所有共享数据均已去标识化，以保护参与者隐私：直接标识符已删除，间接标识符已尽可能减少。眼动追踪数据仅包含数值测量（注视坐标、注视持续时间、时间戳及相关指标）；研究期间未采集或存储参与者的视频录像。您可通过以下链接访问这些数据：https://osf.io/4wjgp/。

CRediT 作者贡献声明

DP：概念构思、调查、方法学、软件、正式分析、数据整理、可视化、撰写初稿、审阅与编辑。FJ：方法学、审阅与编辑。DS：审阅与编辑。FS：资源、监督、审阅与编辑。FM：监督、概念构思、调查、方法学、软件、撰写初稿、审阅与编辑。

致谢

我们谨向 Rita Hansl、Alex Karner、Kathrin Kostorz、Cindy Lor、Daniel Reiter、Annika Trapple、Nicole Wimmer 和张梦凡致谢，感谢他们在开发这项基于 Web 的实验期间所作的贡献。我们还要感谢 Hermann Kaindl 在本研究开发过程中提供的有益讨论和反馈。

利益冲突

作者声明不存在任何利益冲突。

资助

本研究由奥地利研究促进署 (FFG) 资助，项目编号：471030、887474 和 927913。FM 获奥地利科学基金 (FWF) 资助，项目编号为 [10.55776/ESP133]。

参考文献

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M.（2018）。面向可解释、可问责且易于理解的系统的趋势与发展轨迹：一项人机交互研究议程。*2018*年计算机系统人因大会论文集，1–18。https://doi.org/10.1145/3173574.3174156

Afsar, B., Silvennoinen, J., Misitano, G., Ruiz, F., Ruiz, A. B., & Miettinen, K. (2023). 设计实证实验以比较交互式多目标优化方法。《运筹学会期刊》，*74*(11)，2327–2338。https://doi.org/10.1080/01605682.2022.2141145

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). 混合智能研究议程：以协作、自适应、负责任且可解释的人工智能增强人类智力。《计算机》，*53*(8)，18–28。https://doi.org/10.1109/MC.2020.2996587

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). 可解释人工智能（可解释人工智能）：概念、分类法、机遇与迈向负责任人工智能的挑战。《信息融合》，*58*，82–115。https://doi.org/10.1016/j.inffus.2019.12.012

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015).使用*lme4*拟合线性混合效应模型。

Brooks, M., Kay-Lambkin, F., Bowman, J., & Childs, S. (2012). 问题性饮酒来访者的自我关怀。正念，*3*(4)，308–317。https://doi.org/10.1007/s12671-012-0106-5

Burnham, K. P., & Anderson, D. R.（编）.(2004). 模型选择与多模型推断. Springer New York. https://doi.org/10.1007/b97636

Bussone, A., Stumpf, S.,& O'Sullivan, D. (2015). 解释对临床决策支持系统中信任与依赖的作用。*2015*年国际医疗信息学会议，160–169。https://doi.org/10.1109/ICHI.2015.26

Cacchiani, V., Iori, M., Locatelli, A., & Martello, S. (2022年). 背包问题——近期进展概述。第二部分：多重、多维与二次背包问题。《计算机与运筹研究》，*143*, 105693。https://doi.org/10.1016/j.cor.2021.105693

Caprara, A., Kellerer, H., & Pferschy, U.（2000年）。多重子集和问题。《*SIAM*优化杂志》，*11*（2），308–319。https://doi.org/10.1137/S1052623498348481

Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S.（2017）。将计划解释视为模型协调：超越独白式解释。https://doi.org/10.48550/arXiv.1701.08317

Chater, N. (1996). 调和知觉组织中的简洁性与似然原则。《心理学评论》, 103(3), 566–581. https://doi.org/10.1037/0033-295X.103.3.566
Christensen, R. H. B. (2023). 序数数据的 *ordinal*—回归模型. https://CRAN.R-project.org/package=ordinal

Coffman, E. G., Garey, M. R., & Johnson,D. S. (1996). 装箱问题的近似算法：综述。载于 *NP*难问题的近似算法 (pp. 46–93). PWS Publishing Co. https://dl.acm.org/doi/10.5555/241938.241940

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). 算法导论（第三版）。麻省理工学院出版社。

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). 算法厌恶：人们在看到算法犯错后会错误地避开算法。《实验心理学杂志：综合》， *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Doshi-Velez, F., & Kim, B. (2017). 迈向可解释机器学习的严谨科学。https://doi.org/10.48550/arXiv.1702.08608

Dumnić, S., Dupljanin, D., Božović, V., & Ćulibrk, D. (2019). PathGame：众包限时条件下人类对旅行商问题的解。《计算智能与神经科学》,*2019*, 1–9. https://doi.org/10.1155/2019/2351591

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). 超越眼动注视：眼动追踪还能揭示认知与认知发展的哪些方面? 《发展认知神经科学》， *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Ehrgott, M. (2005). 多准则优化 （第二版）。施普林格。

Feldman, J. (2016). 知觉与认知中的简约原则。《*WIREs*认知科学》， *7*(5)，330–340. https://doi.org/10.1002/wcs.1406

Fox, M., Long, D., & Magazzeni, D. (2017). 可解释规划。https://doi.org/10.48550/arXiv.1709.10256

Franco, J. P., Doroc, K., Yadav, N., Bossaerts, P., & Murawski, C. (2022年). 计算难度的任务无关指标可预测人类认知表现。《style id='1'>《科学报告》, *12*(1), 12914. https://doi.org/10.1038/s41598-022-16565-w

Franco, J. P., Yadav, N., Bossaerts, P., & Murawski, C. (2021). 计算任务的一般属性可预测人类努力和performance。数学心理学杂志,*104*, 102592. https://doi.org/10.1016/j.jmp.2021.102592

Gigerenzer, G., & Gaissmaier, W. (2011). 启发式决策。《心理学年度评论》， *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gollan, B., & Raggam, P. (2025). 超越注视：通过一种创新的眼动追踪生物标志物量化意识知觉。*ACM*人机交互论文集， *9*(3), ETRA06:1–ETRA06:17. https://doi.org/10.1145/3725831

Gunawan, A., Kendall, G., Lee, L. S., McCollum, B., & Seow, H.-V. (2021). 多学科调度趋势。《运筹学会期刊》， *72*(8), 1689–1690. https://doi.org/10.1080/01605682.2021.1947755

Gurski,F., Rehs, C., & Rethmann, J. (2019). 背包问题：参数化视角。《理论计算机科学》, *775*, 93–108. https://doi.org/10.1016/j.tcs.2018.12.019

Helmholtz, H. L. F. von. (1962). 《生理光学论著》. Dover. (原著发表于 1909)

Heppner, P. P., & Petersen, C. H. (1982). 个人问题解决量表的发展及其意义。《咨询心理学杂志》, *29*(1), 66–75. https://doi.org/10.1037/0022-0167.29.1.66

Ibs, I., Ott, C., Jäkel, F.,& Rothkopf, C. A. (2024). 从人类解释到可解释人工智能：约束优化的启示。《认知系统研究》, 101297. https://doi.org/10.1016/j.cogsys.2024.101297

Ibs,I., & Rothkopf, C. A. (2026). 基于人类解释为约束优化生成理由。载于 R. Guidotti, U. Schmid, & L. Longo (编),可解释人工智能（第 162–184 页，第 2576 卷）。施普林格·自然瑞士。https://doi.org/10.1007/978-3-032-08317-3_8

Johnson, D. S., Demers, A., Ullman, J. D., Garey, M. R., & Graham, R. L. (1974年). 简单一维装箱算法的最坏情况性能界。《 *SIAM* 计算杂志》， *3*(4), 299–325. https://doi.org/10.1137/0203025

Kahneman, D. (2011).《思考，快与慢》（第1版）。Farrar, Straus and Giroux。

Kellerer, H., Pferschy, U., & Pisinger, D. (2004).背包问题。Springer Berlin Heidelberg。https://doi.org/10.1007/978-3-540-24777-7

Krakowski, S., Luger, J., & Raisch, S. (2023). 人工智能与竞争优势来源的变化。《战略管理杂志》， *44*(6)，1425–1452。https://doi.org/10.1002/smj.3387

Kyritsis, M., Gulliver, S. R., Feredoes, E., & Stouraitis, V. (2022). 人们对欧几里得旅行商问题中相互竞争的解的最优性的感知。《认知系统研究》， *74*，1–17。https://doi.org/10.1016/j.cogsys.2022.02.001

Lee, J. D. 与 See, K. A.（2004）。对自动化的信任：为适当依赖而设计。《人为因素》， *46*(1)，50–80。https://doi.org/10.1518/hfes.46.1.50_30392

Lipton, Z. C.（2017）。模型可解释性之谜。https://doi.org/10.48550/arXiv.1606.03490

Loaiza-Ganem, G. 与 Cunningham, J. P.（2019）。连续伯努利分布：修正变分自编码器中的普遍错误。https://doi.org/10.48550/arXiv.1907.06845

Luce, R. D.（1986）。反应时：其在推断基本心理组织中的作用。牛津大学出版社；克拉伦登出版社。

Lüdecke, D.、Ben-Shachar, M. S.、Patil, I.、Waggoner, P. 与 Makowski, D.（2021）。*performance*：用于评估、比较和检验统计模型的 *R*软件包。

MacGregor, J. N., & Chu, Y. (2011). 人类在旅行商及相关问题上的 performance：综述。《问题求解杂志》， *3*(2). https://doi.org/10.7771/1932-6246.1090

Marzouk, M., & Kamoun, H. (2021). 通过类比装箱问题进行护士与患者分配：突尼斯医院案例。《运筹学会期刊》，*72*(8), 1808–1821. https://doi.org/10.1080/01605682.2020.1727300

Miller, T.（2019）。人工智能中的解释：来自社会科学的洞见。人工智能，*267*，1–38。https://doi.org/10.1016/j.artint.2018.07.007

Murawski, C.，& Bossaerts, P.（2016）。人类如何解决复杂问题：背包问题案例。《科学报告》，*6*(1)，34851。https://doi.org/10.1038/srep34851

Nakagawa, S.，& Schielzeth, H.（2013）。从广义线性混合效应模型中获得 R2 的通用简易方法。《生态学与进化方法》，*4*(2)，133–142。https://doi.org/10.1111/j.2041-210x.2012.00261.x

Narayanan, M.、Chen, E.、He, J.、Kim, B.、Gershman, S.，& Doshi-Velez, F.（2018）。人类如何理解来自机器学习系统的解释？对解释的人类可解释性进行评估。https://doi.org/10.48550/arXiv.1802.00682

Ott, C., & Jäkel, F.（2023）。*SimplifEx*：简化并解释线性规划。https://osf.io/v4xmc/

Palan, S., & Schitter, C.（2018）。Prolific.ac——用于在线实验的被试样本库。《行为与实验金融学期刊》，*17*，22–27。https://doi.org/10.1016/j.jbef.2017.12.004

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J.（2016）。Webgazer：利用用户交互实现可扩展的网络摄像头眼动追踪。第二十五届人工智能国际联合会议论文集，3839–3845。

Perron, L., & Didier, F.（2024年5月7日）。*CP-SAT*（版本v9.11）。https://developers.google.com/optimization/cp/cp_solver/

R 核心团队。（2023）。*R*：一种用于统计计算的语言和环境。维也纳，奥地利。https://www.R-project.org/

Rudin, C.（2019）。不要再为高风险决策解释黑箱机器学习模型，而应改用可解释模型。《自然·机器智能》，*1*(5)，206–215。https://doi.org/10.1038/s42256-019-0048-x

Sweller, J. (1988). 问题求解过程中的认知负荷：对学习的影响。《认知科学》，*12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Thompson, B., & Griffiths, T. L. (2021). 人类偏见限制累积创新。《英国皇家学会学报*B*辑：生物科学》, *288*(1946), 20202752. https://doi.org/10.1098/rspb.2020.2752

Tversky, A., & Kahneman, D. (1974). 不确定性下的判断：启发式方法与偏见。《科学》，*185*(4157), 1124–1131. 检索于 2025 年 2 月 14 日，网址：https://www.jstor.org/stable/1738360

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). 动画：能否发挥促进作用?《国际人机研究杂志》, *57*(4), 247–262. https://doi.org/10.1006/ijhc.2002.1017

van der Helm, P. A.（2000年）。视觉感知中的简洁性与似然性：从惊奇度到精确度。《心理学通报》，*126*(5)，770–800。https://doi.org/10.1037/0033-2909.126.5.770

Zerilli, J., Bhatt, U., & Weller, A.（2022年）。透明度如何调节人们对人工智能的信任。《*Patterns*》，*3*（4），100455。https://doi.org/10.1016/j.patter.2022.100455

附录A

补充方法

模型拟合与比较

通用约定

- 软件和估计： `clmm`（`thresholds = "symmetric"`， `link = "logit"`，拉普拉斯近似）； `lmer` 使用 `REML = FALSE`。所有预测变量均在完整样本范围内进行 z 标准化。

- 候选模型生成：仅包含二阶交互作用，且仅限于复杂度预测变量与调节变量之间。不包含主效应之间的交互作用。最大无序度绝不会与视觉顺序复杂度同时纳入同一模型。预测变量可以单独纳入；调节变量仅与其对应的主效应一同纳入。

- 随机效应（在每项分析中保持一致）：所有模型均包含参与者随机截距；在收敛允许的情况下，为纳入的复杂度主效应（以及存在时的对角线差异度）设置随机斜率。交互作用不设置随机斜率。如果最大结构无法收敛，则在所有候选模型中统一简化，直至收敛。

- 模型选择：采用基于 AIC 的选择；保留 AIC 比所有竞争模型至少低 2 个单位的模型（$\Delta\text{AIC} > 2$）。如果多个模型的 AIC 相差不超过 2 个单位，则视为等效（Burnham 和 Anderson，2004），并选择最简单的模型（参数最少）。

- 预处理：选择和注视的复杂度预测变量（以及对角线差异度）采用带符号的右侧–左侧差值；反应时 采用绝对差值。调节变量不计算差值。

选择（有序结果）

- 结果变量：选择（右侧与左侧； `clmm`）。

- 主效应： Δ启发式相关复杂度， Δ组合复杂度， Δ视觉顺序复杂度， Δ对角线差异。

- 调节变量：PD、PSI、PSE、MD、HO。
- 交互作用： ΔHC/Δ组合复杂度/Δ视觉顺序复杂度×调节变量。
- 约束：不要将 MD 与 Δ视觉顺序复杂度纳入其中。

反应时间（连续型结果）

- 结果：反应时对数（`lmer; REML = FALSE`）。
- 主效应： |ΔHC|、 |Δ组合复杂度|、 |Δ视觉顺序复杂度|、 |Δ对角线差异|。
- 其他协变量：PD、PSI、PSE、HO、MD。
- 无调节变量/交互作用
- 约束：不要将 MD 与 |Δ视觉顺序复杂度| 同时纳入。

注视偏向（二项型结果）

- 结果： `cbind(R, L)`, 二项分布（逻辑斯蒂）， `glmer`。
- 主效应： Δ启发式相关复杂度、 Δ组合复杂度、 Δ视觉顺序复杂度、 Δ对角线差异。
- 调节变量：PD、PSI、PSE、MD、HO。
- 交互作用： Δ启发式相关复杂度/Δ组合复杂度/Δ视觉顺序复杂度×调节变量。
- 约束：不要将 MD 与 Δ视觉顺序复杂度一同纳入。

刺激与试次生成

问题和解实例的生成

问题实例通过包含 20,000 次迭代的仿真过程生成。在每次模拟迭代中，通过随机确定物品和箱的数量来构造一个装箱问题实例。物品数量从离散均匀分布 U(7, 9) 中抽取，箱的数量则从

U(4, 6) 中抽取。物品尺寸从步长为 5、范围为 5 到 100 的离散均匀分布中抽取。总体箱子容量基于从 [0.8, 1.0] 中均匀抽取的负载容量比确定。随后，我们将总体箱子容量分配到各个箱中，同时确保箱子容量最小为 10、最大为 100，且以 10 为步长。为确保问题实例有意义，我们对每个实例进行筛选，确保没有物品尺寸超过最大箱子容量，也没有箱子容量小于最小物品尺寸。

生成的问题实例包含一个物品负载向量和一个相应的箱子容量向量。使用约束规划可满足性（CP-SAT）求解器（Perron & Didier，2024）计算每个问题实例的最优解。由于求解器默认只返回一个最优解，我们通过迭代添加约束来排除先前找到的解，直到获得最多 100 个不同的最优解。当找到 100 个解，或新解的目标值低于当前最优解的目标值时，过程便会停止。

确定最优解后，我们移除了冗余解，具体做法是检查箱体组成是否等价；例如，如果两个解都包含由相同物品组成的箱子，那么这两个解就被视为等价。比如，一个箱子包含尺寸为 100、90、80、40 的物品，另一个箱子包含尺寸为 100、80、90、40 的物品；这确保了两个解中的所有箱子在内容上都相匹配。最后，我们剔除了只有一个独特最优解的问题实例，以便比较同一问题实例的多个解。此外，我们还计算了针对解的特定指标，包括 HC 和 CC，二者均在正文中有所介绍。

应用所有筛选条件后，这 20,000 次模拟共得到 13,269 个问题实例。

试次的生成

后续的试次生成程序旨在为实验的两个不同部分生成试次：问题解决试次和评估试次。对于

问题解决试次，我们根据负载容量比（表示问题难度的指标），从所有先前生成的问题实例中通过分位数抽样选取了七个问题实例。具体而言，七个分位数按 0 到 1 的范围等距确定，并为每个分位数选取相应的问题实例。随后，为这七个问题解决试次各配上一个任意的最优解（即 CP-SAT 求解器针对每个问题实例生成的第一个解），以便为参与者提供清晰的示例。这七个试次只需生成一次，供实验中的所有参与者使用。

针对每位参与者，均从生成的问题实例及其对应的解实例全集中，为其单独选取一组包含 25 个评估试次的试题。这些试次通过在每个问题实例内配对解实例构建而成。为确保在关键复杂度指标上进行均衡的分层抽样，首先根据负荷-容量比的分位数，将问题实例划分为不同的难度等级（低、中、高）。对于每项关注的指标（即 HC 的绝对差和 CC 的绝对差），通过针对每个难度等级分别选取该指标位于前 10% 的两个解对，得到六个极端化评估试次。此外，按照问题难度和问题规模进行分层抽样，生成三个随机评估试次。另有三个试次通过克隆先前生成的随机评估试次并施加视觉变换（调换物品和箱子的顺序）得到：其中一对只对第一个解施加视觉变换，另一对只对第二个解施加视觉变换，还有一对则对两个解都施加视觉变换（在代码库中称为“重复随机试次”）。通过按问题规模进行分层抽样，另外安排了两个随机试次。与前述试次不同，每对试次包含两个相同的解，但视觉呈现有所不同：第一对只对第二个解施加视觉变换，第二对只对第一个解施加视觉变换（在代码库中称为“相同解随机试次”）。

这 20 个试次针对每位参与者分别抽样。此外，评估试次中还加入了两个注意力检测试次和三个一致性试次，以评估参与者的专注程度和一致性。这五个试次对所有参与者均相同，且只抽样一次。对于前者，选取了两个难度较低且包含相同解的问题实例；在实验中，对这些实例的任何偏好都表明参与者未认真投入任务。为生成一致性试次，我们首先筛选问题集，找出问题难度处于中等水平且至少有三个最优解的案例，因为一致性试次要求同一问题具有三个不同的解。为确保解的多样性，我们通过计算所选指标——组合复杂度的取值范围（即最大值减去最小值）来评估每个候选问题。此外，我们还通过计算该指标分布的中位数与均值之差的绝对值，确定其不对称性得分。该得分有助于保持指标值分布的对称性，从而确保解的差异性均衡且具有代表性。为确保指标具有充分的变异性，我们保留了取值范围达到或超过第95百分位数的问题。在这一筛选后的集合中，我们选取了不对称性得分最低的问题。针对选定的问题，选取组合复杂度分别为最小值、中位数和最大值的代表性解，并将其配对，形成最终的一致性试次。

整个评估试次序列的构建方式是：为一致性试次和注意力检测试次分配特定的试次位置，同时随机安排其余评估试次。实验设计总共包含七个问题解决试次，以及一组结构化的评估试次，其中包括极端化试次、随机试次（有和无视觉操控）、一致性试次和注意力检测试次。通过这种谨慎的分层，确保各试次在解实例复杂度和问题难度方面保持平衡，从而尽可能减少潜在的抽样偏差。

近似对角指派矩阵的构建

用于计算 DD（对角线差异）的近似对角线是一个纯理论上的分配矩阵，用来控制 HC（启发式相关复杂度）；它模拟一条直线对角线，并非旨在构成现实或最优的解。它唯一的用途是提供一个简单的几何基准，用于衡量某个解偏离对角线模式的程度。下面是创建该矩阵的 Python 代码：

```
def create_diagonal_matrix_approximated(n_rows, n_cols):
    # 1. Create an all-zero matrix M of size rows x cols.
    M = [[0 for _ in range(n_cols)] for _ in range(n_rows)]

    # 2. For each row i = 0...rows - 1
    for i in range(n_rows):
        # a. col_index <- round[i * (cols - 1) / (rows - 1)]
        col_index = round(i * (n_cols - 1) / (n_rows - 1))

        # b. Set M[i, col_index] <- 1
        M[i][col_index] = 1

    # 3. Return M.
    return M
```

步骤 2a 中的插值将“1”条目均匀地从左上角分布到右下角，使行 $\neq$ 列呈现阶梯状对角线。

组合复杂度模型的校准

探索性实验的校准

组合复杂度（CC）的混合模型在数据收集之前完成了校准，原因在于：（a）试次生成程序依赖于该模型为每个解实例赋予的复杂度值；（b）后续的混合效应分析要求为每个刺激材料设定一个单一且固定的复杂度估计值。校准所使用的语料库包含145个此前获得的最优解。为了在随机初始化之外为校准建立一个有原则的目标，我们计算了三个与CC直觉相一致的简单指标，在整个语料库中对其进行z标准化，

随后进行了主成分分析（PCA），并保留了第一主成分（方差占比最大），将其作为单变量综合得分，充当经验基准。

对于每个解，设 $x \in \{0,1\}^{n\times m}$ 为分配矩阵（物品 $\times$ 箱），$z = (z_1, \ldots, z_n)$为物品尺寸向量，$w = (w_1, \ldots, w_m)$为箱容量向量；对于箱i，将分配给它的物品集合记为 $J_i = \{ j \mid x_{ij} = 1\}$。

- **分配方差 (AV)**：各箱分配物品数量的标准差，

$$\mathrm{AV} \;=\; \mathrm{sd}_i\left(\sum_{j=1}^{n} x_{ij}\right). \tag{A1}$$

- **平均差值 (AD)**：对于至少分配了一个物品的箱，最大已分配物品放入后剩余容量的均值，

$$\mathrm{AD} \;=\; \mathrm{mean}_{i\in I^*}\left(w_i \;-\; \max_{j\in J_i} z_j\right), \quad I^* \;=\; \left\{ i \,\middle|\, \sum_{j=1}^{n} x_{ij} > 0 \right\}. \tag{A2}$$

- **平均比率 (AR)**：对于至少分配了一个物品的箱，定义

  r_i =均值$_j(z_j x_{ij})/w_i$ （该均值$_j$ 包含未分配物品对应的零值），然后

$$\mathrm{AR} \;=\; \frac{1}{\mathrm{mean}_{i\in I^*}\left(r_i\right)}. \tag{A3}$$

CC 模型包含三个连续参数（箱中物品数量的惩罚项 p、空余空间混合分布的尺度 σ、狄利克雷浓度 α）和两个类别开关（空余空间分布：正态分布、拉普拉斯分布或连续伯努利分布；可选的狄利克雷校正），共形成六种模型变体。对于每种变体，我们使用有界拟牛顿优化（L-BFGS-B），在约束条件下拟合了 $(p,\ \sigma,\ \alpha)$$p \in (0,\ 1)$、$\sigma > 0$ （适用于正态分布和拉普拉斯分布变体；对于连续伯努利分布，

$\sigma \in (0,\ 1))$、 $\alpha > 0$，起始值设为 $p = 0.50$、 $\sigma = 0.05$、 $\alpha = 2.00$。损失函数为模型预测的复杂度与主成分分析综合得分之间的皮尔逊相关的负值，因此最小化损失即可最大化二者的一致性。我们保留最终相关性最高的参数集，用于探索性研究（试次生成和分析）。

搜索确定，带有狄利克雷校正的截断正态空空间为最优，其参数估计值为 $p = 0.977$（几何），$\sigma = 0.103$（标准差），以及 $\alpha = 1.620$（狄利克雷浓度）。

验证性实验的校准

探索性研究后重新调整了组合复杂度模型，原因是：(a) 试次生成程序依赖于该模型为每个解实例指定的复杂度值；(b) 在混合效应分析中将其用作预测变量，需要采用一套固定的参数设定。校准依据探索性样本（n $= 73$ 名参与者，1,664 个试次）表达的左右偏好进行。对于每个试次，模型都为两个备选解分别生成一个复杂度估计值；二者的差值作为序数选择变量的唯一预测变量（明确偏左、稍微偏左、稍微偏右、明确偏右）。与探索性校准一样，对三个连续参数进行了优化：箱内物品数量的几何惩罚、未分配空间的权重，以及熵项的缩放因子。此外，还交叉考察了两个类别开关：空分配所采用的分布（正态分布、拉普拉斯分布或连续伯努利分布），以及可选的狄利克雷校正，由此得到六种离散模型变体。对于每种变体，均使用有界优化拟合三个连续参数，起始值分别为 0.50、0.05 和 2.00。损失函数为四级观察选择与模型隐含概率之间的序数对数损失；最小化该值即可最大化预测准确率。搜索结果表明，采用连续伯努利分布描述空空间并进行狄利克雷校正的模型最优，参数估计值为 $p = 0.043$, $\sigma =$

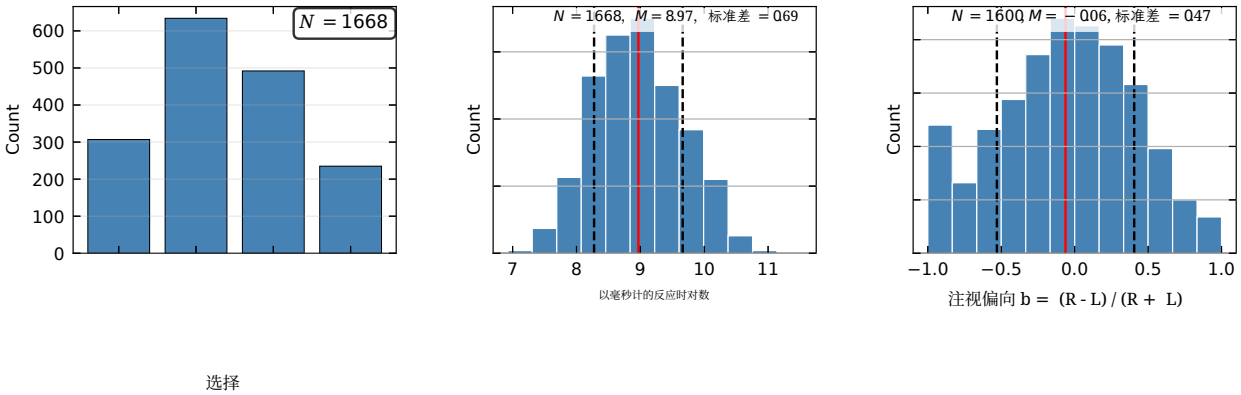
0.426, 和 $\alpha = 0.984$。该配置用于生成最终试验集，并在所有后续分析中保持固定。

附录B

补充结果

图B1

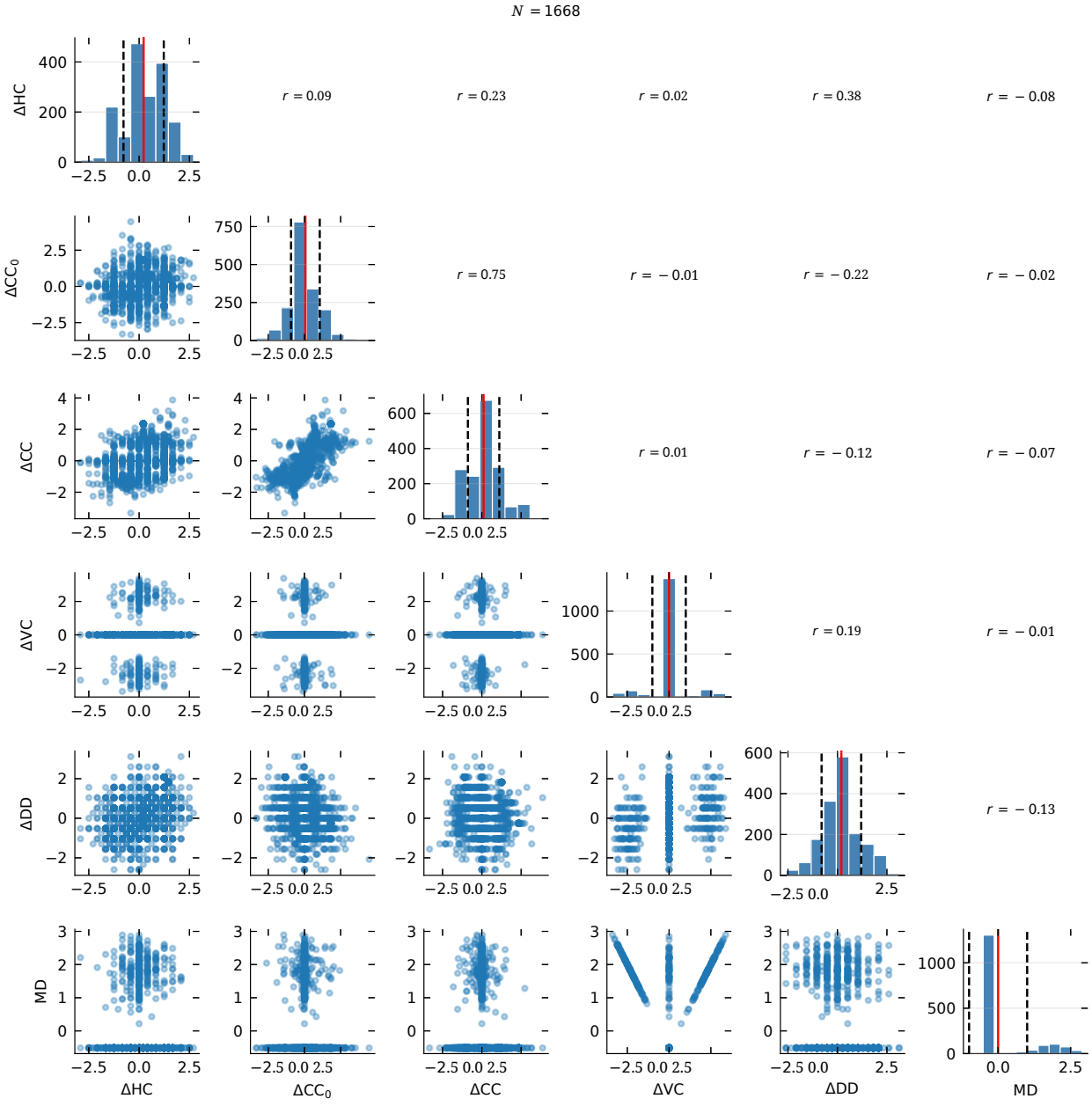
因变量的行为分布



注：直方图展示了评估试次中三个因变量的分布：选择（四个有序类别）、对数反应时（反应时对数）和注视偏向 $b = (R - L)/(R + L)$。

图 B2

选择与注视的解对层级预测变量散点矩阵

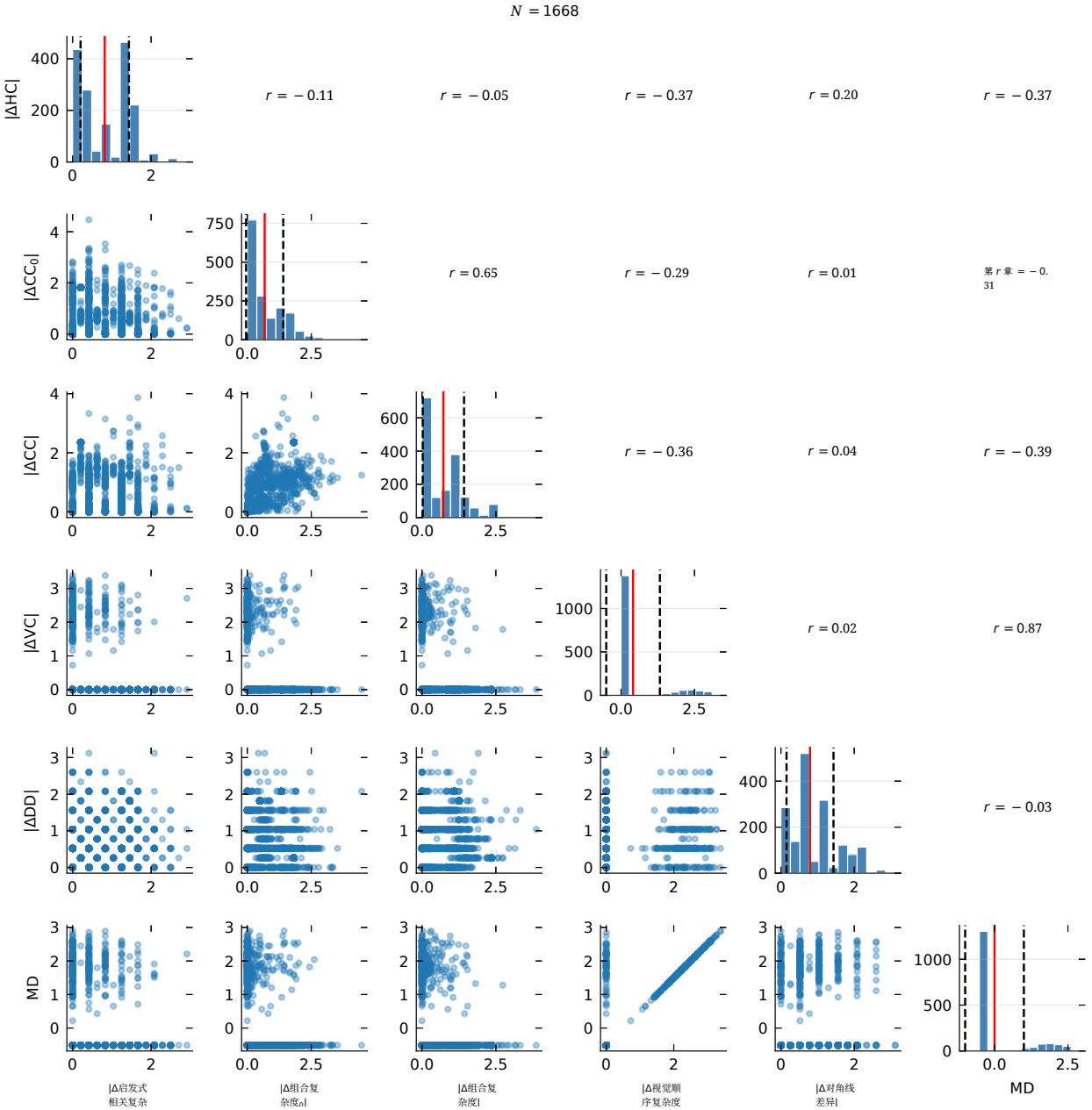


注。带符号的标准 化差值，用于选择/注视分析（ΔHC、 ΔCC、 ΔVC、 ΔDD）以及

最大无序度（MD）。CC_0 对应于探索性研究中使用的未校准CC。上三角显示Pearson的 r；对角线显示分布（均值为红色，标准差为虚线）。

图 B3

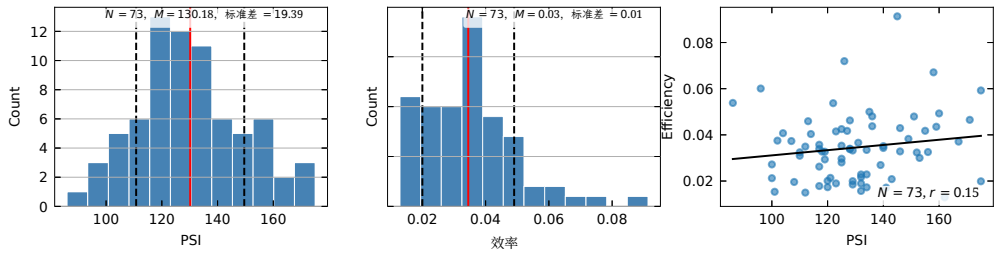
反应时间的解对层面预测变量散点矩阵



注。绝对值　　RT分析中使用的标准化差值（|Δ|HC、 |Δ|CC、 |Δ|VC、 |Δ|DD）和最大无序度（MD）。CC_0 对应探索性研究中使用的未校准CC。上三角显示皮尔逊 r；对角线显示分布（红色表示均值，虚线表示标准差）。

图B4

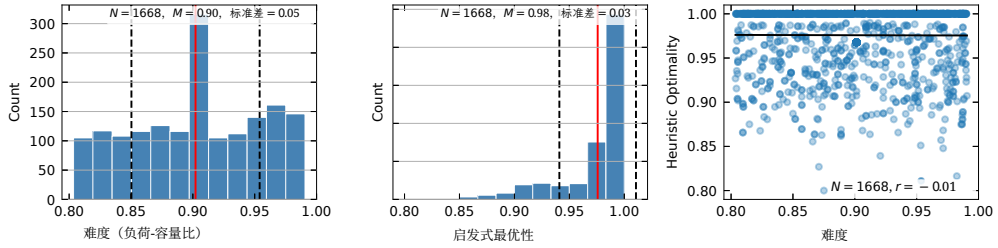
参与者层面变量



注。参与者层面调节变量（PSI 总分、问题解决效率）的分布与相关（Pearson）。汇总统计量见正文；图示记录了调节分析中所用的取值范围与关联强度。

图B5

问题层面变量



注。问题层面调节变量（难度：负荷-容量比；启发式最优性）的分布与相关（Pearson）。为体现透明度，展示调节检验中的取值范围及潜在混杂。

附录C

探索性研究

参与者

鉴于本研究所考察的新颖范式，研究采用了序贯数据收集方法，反复评估结果并优化样本量。最初招募了3名参与者，以对研究程序进行初步审查，并处理任何即时的方法学问题。在确认可行性后，后续队列分别纳入了6名、12名、24名、36名和36名参与者，最终总样本量达到114名。这种逐步增加的方式有助于持续评估数据变异性，并及早发现潜在效应。最终样本量根据所观察到的关键发现趋于饱和的情况确定，同时着重捕捉多样化的参与者反应，以增强探索性洞见的稳健性。在整个研究过程中，这种方法对于保持灵活性和优化资源利用发挥了重要作用。

共有114名从 Prolific 招募的参与者完成了该研究，排除后剩余73名参与者和1,664个评估试次。年龄范围为19至64岁（M $= 36.49$，SD $= 11.06$），样本中男性占58.90%，女性占41.10%。参与者完成实验的中位时间为29.73分钟（第25–75百分位数 $= 21.73$–40.80分钟）。

组合复杂度（CC）的参数

在每项实验（探索性或验证性）之前，均实施了专门的校准程序以确定这些参数（详见附录A）。在我们的探索性分析中，该程序得到以下参数：空白空间采用截断正态分布，并施加狄利克雷校正，尺度参数为 $\sigma = 0.103$、$p = 0.977$和 $\alpha = 1.620$。

结果

选择

四项预测变量均为有序结果的显著负向预测变量。差值每增加一个标准差，选择更复杂解的优势分别降低：启发式相关复杂度降低33%（优势比 $= 0.67$，95%置信区间 $[0.59, 0.77]$），组合复杂度降低14%（优势比 $= 0.86$，95%置信区间 $[0.78, 0.95]$），视觉顺序复杂度降低41%（优势比 $= 0.59$，95%置信区间 $[0.51, 0.68]$），对角线差异降低19%（优势比$= 0.81$，95%置信区间 $[0.71, 0.91]$），且存在相当大的参与者间变异（表C1）。图 C1展示了刺激复杂度的差值如何反映在参与者的选择行为中。条件Nakagawa R^2 为0.29表明，模型的固定效应和随机效应解释了数据中约29%的方差，而边际NakagawaR^2 为0.13表明，仅固定效应约占13%。该值远低于0.817的连贯性上限；该上限源自一致性试次中表现出传递性判断的81.69%参与者（58名参与者的判断为传递的，13名参与者的判断为非传递性判断），这表明仍有相当大的决策方差未被当前这组结构性预测变量所捕捉。

在评估试次中，选择比例分别为：明确偏左 22.7%、稍微偏左 31.3%、稍微偏右 28.7%、明确偏右 17.3%。核心预测变量（带符号差值）之间的两两相关较弱（最大值 $|r| = 0.33$）。

表 C1预测选择的序数混合效应模型（探索性研究）

Term	估计值	SE	z	p	置信区间下限	置信区间上限
固定效应						
中心阈值	0.123	0.066	1.871	0.061	-0.006	0.252
阈值间距	1.681	0.050	33.533	$<$ 0.001	1.583	1.779
ΔHC	-0.401	0.069	-5.824	$<$ 0.001	-0.536	-0.266
ΔCC	-0.151	0.050	-2.986	0.003	-0.250	-0.052
ΔVC	-0.530	0.073	-7.266	$<$ 0.001	-0.673	-0.387
ΔDD	-0.216	0.063	-3.444	$<$ 0.001	-0.339	-0.093
随机效应						
标准差（截距 | 受试者）	0.381					
标准差（Δ启发式相关复杂度 | 受试者）	0.371					
标准差（Δ组合复杂度 | 受试者）	0.261					
标准差（Δ视觉顺序复杂度 | 受试者）	0.440					
标准差（Δ对角线差异 | 受试者）	0.299					

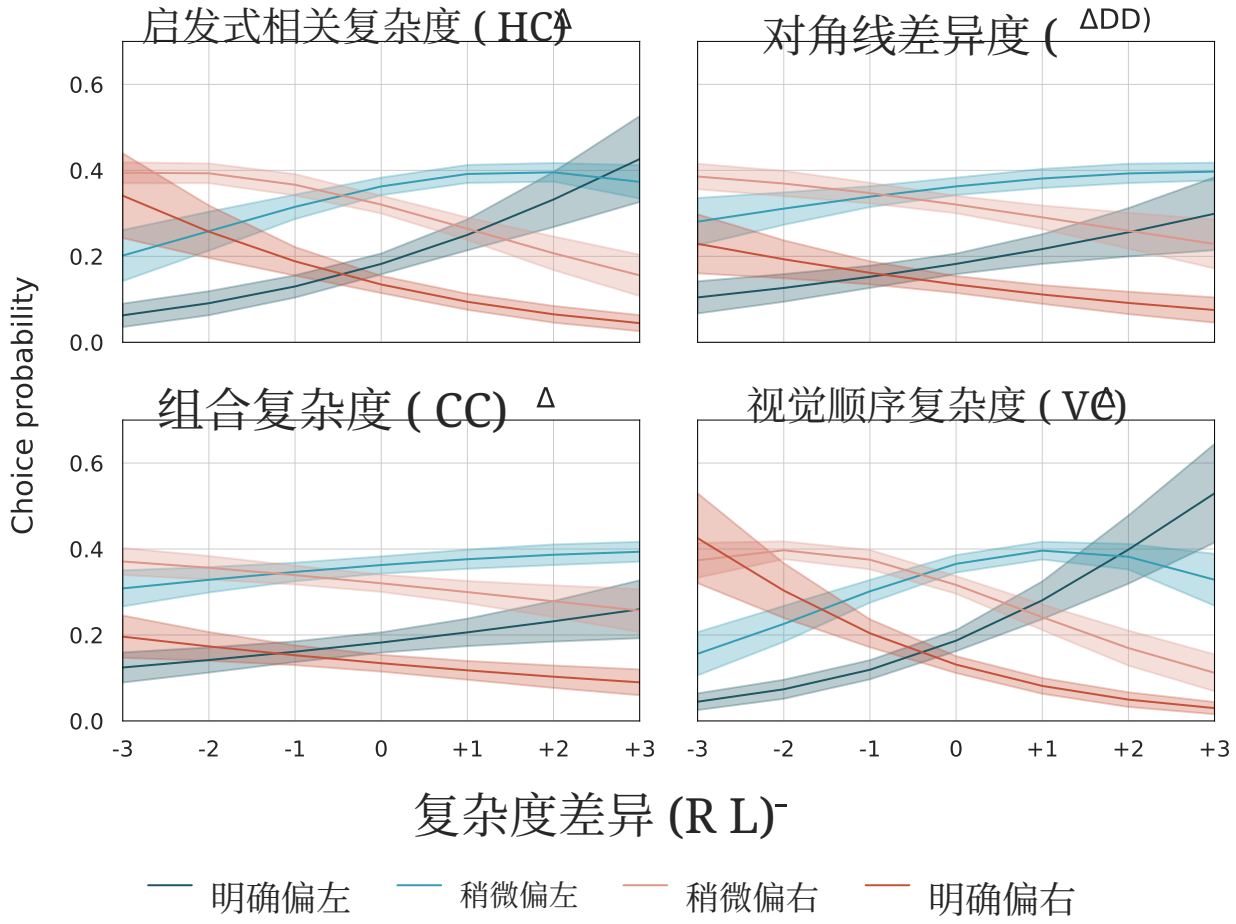
注。选择 $\sim$ Δ启发式相关复杂度 $+$ Δ组合复杂度 $+$ Δ视觉顺序复杂度 $+$ Δ对角线差异 $+$ （1 $+$ Δ启发式相关复杂度 $+$ Δ组合复杂度 $+$ Δ视觉顺序复杂度 $+$ Δ对角线差异| 受试者）。 N（obs） $= 1664$， N（$subj$） $= 73$，对数似然 $=$-2127.5。预测变量每变化 1 个标准差，其优势比（OR）为 OR $=$exp(β)；相应的优势百分比变化为 100 $\cdot$(exp(β) $- 1$)。

反应时间

HC、CC 和 VC 的绝对差值越大，选择速度越快；而 DD 协变量对反应时间没有显著影响（见表 C2）。模型的边际 R^2为 0.004，条件 R^2 为 0.604，表明固定效应解释的方差较少，而被试差异的随机效应则解释了大量额外方差。相应的原始反应时中位数为 10812 毫秒（第 25–75 百分位数为 $=$ 6320–18378 毫秒）。

图 C1

复杂度差异与预测选择概率的关系（探索性研究）



注。面板对应三种复杂度指标（HC = 启发式相关复杂度、CC =组合复杂度、VC = 视觉顺序复杂度）以及协变量对角线差异度（DD）。复杂度差异经过标准化（0 = 无差值； 1 = 一个标准差）。彩色线条表示四种行为反应（“明确偏左”“稍微偏左”“稍微偏右”“明确偏右”）的模型预测概率；阴影带表示 95% 置信区间（CIs）。

表 C2预测反应时的线性混合效应模型

Term	估计值	SE	t	p	置信区间下限	置信区间上限
固定效应						
截距	9.399	0.075	124.644	< 0.001	9.250	9.549
|Δ启发式相关复杂度|	-0.067	0.021	-3.247	0.001	-0.107	-0.027
|Δ组合复杂度|	-0.041	0.016	-2.552	0.011	-0.072	-0.009
|Δ视觉顺序复杂度|	-0.051	0.015	-3.324	< 0.001	-0.080	-0.021
|Δ对角线差异|	-0.009	0.019	-0.454	0.650	-0.047	0.029
随机效应						
标准差（截距）	0.586					
标准差（残差）	0.476					

注：反应时 $\sim$ |Δ启发式相关复杂度| + |Δ组合复杂度| + |Δ视觉顺序复杂度| + |Δ对角线差异| + (1 | 受试者)。结果变量：对数反应时（反应时对数）。 $N(obs) = 1664$, $N(subj) = 73$, 对数似然 =-1256.5。

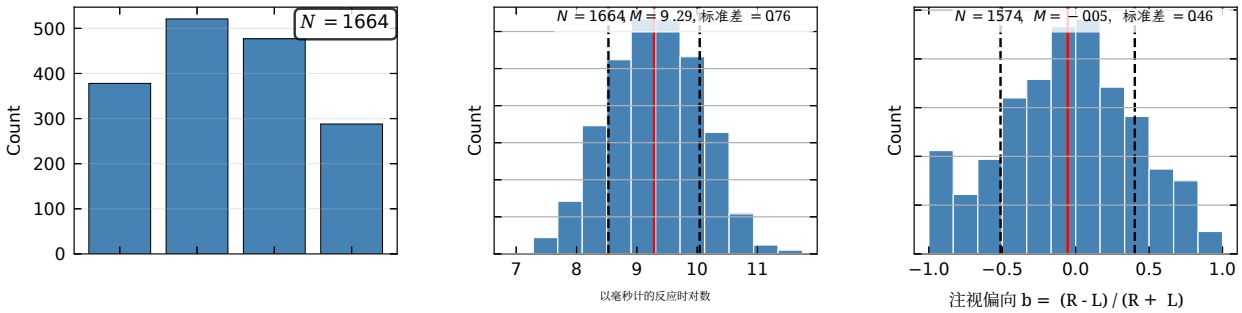
注视偏向

根据基于AIC的模型选择程序，最终选择了仅截距模型。截距为 -0.385（标准误 $= 0.106$，95% 置信区间 [-0.594, -0.177], $p = < 0.001$）。模型拟合指标较低（边际 $R^2 = 0.000$；条件 $R^2 = 0.166$），这与仅截距模型的结果以及复杂度对侧向注视时长没有可靠影响的结论一致。注视偏向（b）的平均值为 -0.055（标准差 $= 0.457$）；没有可用注视数据的试次占 5.4%。分析基于 69 名参与者的 1,574 个观测值。这些结果表明，在探索性研究中，没有证据显示解之间的复杂度或DD左右差异会影响注视。

描述性统计

图 C2

因变量的行为分布（探索性研究）

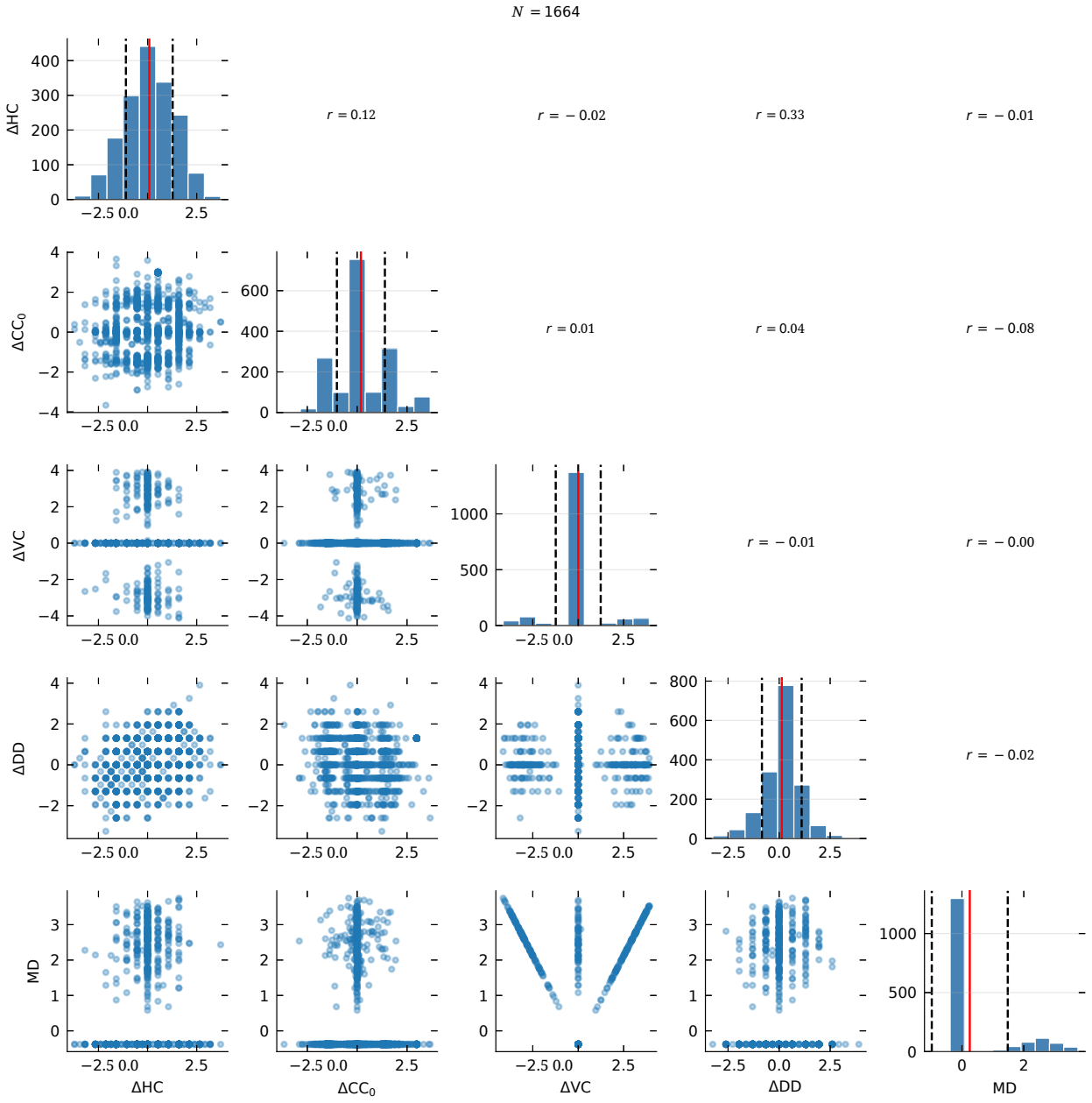


注。直方图显示了三个因变量在评估试次中的分布：选择

(fo（有序类别）、对数反应时（反应时对数）和注视偏向 $b = (R - L)/(R + L)$.

图 C3

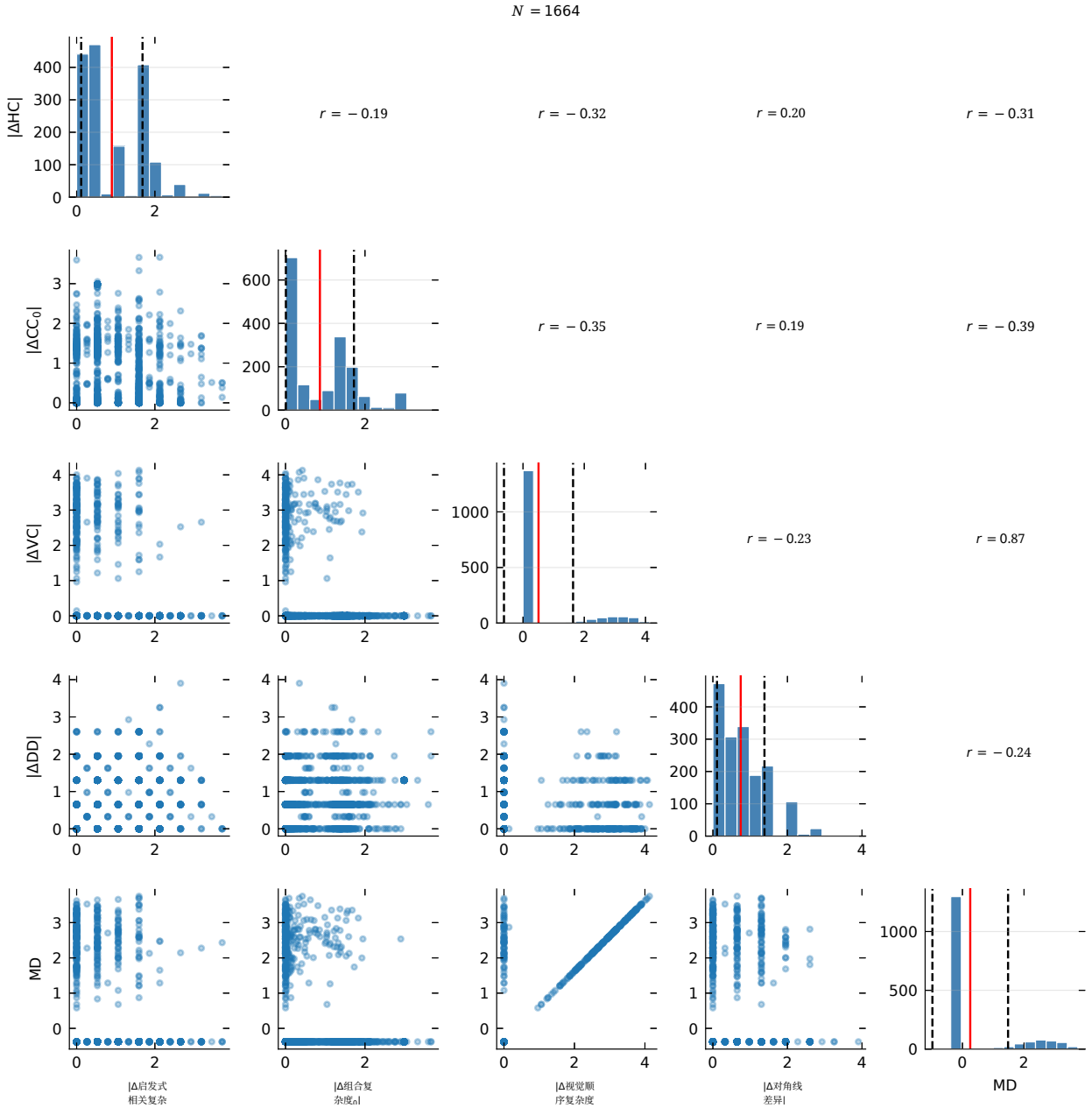
选择与注视的解对层级预测变量散点矩阵（探索性研究）



注。选择/注视分析中使用的带符号的标准化差异（Δ启发式相关复杂度、 Δ组合复杂度、 Δ视觉顺序复杂度、 Δ对角线差异）以及最大无序度（MD）。CC_0 对应于探索性研究中使用的未校准CC。上三角显示皮尔逊 r；对角线显示分布（红色为均值，虚线为标准差）。

图 C4

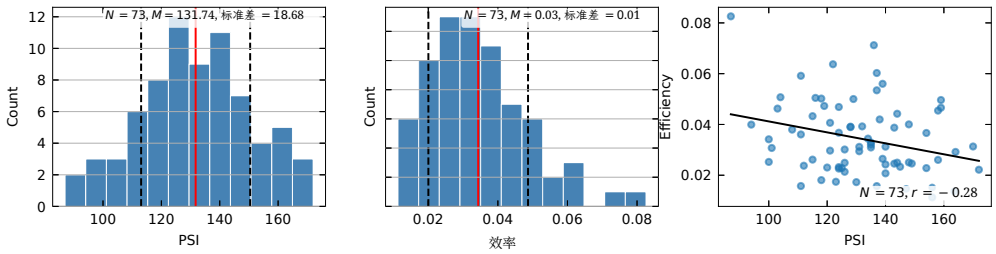
探索性研究中反应时间的解对层面预测变量散点矩阵



注。RT分析中使用的绝对标准化差异（|Δ|启发式相关复杂度、 |Δ|组合复杂度、 |Δ|视觉顺序复杂度、 |Δ|对角线差异）以及最大无序度（MD）。CC_0 对应于探索性研究中使用的未校准CC。上三角显示皮尔逊 r；对角线显示分布（红色为均值，虚线为标准差）。

图 C5

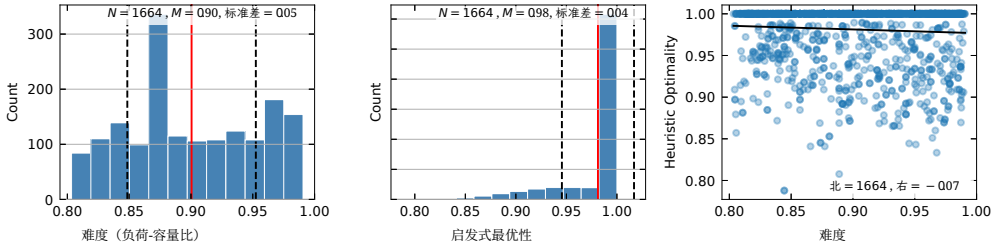
参与者层面变量（探索性研究）



注。参与者层面调节变量（PSI 总分、问题解决效率）的分布及相关性（皮尔逊相关）。汇总统计量见正文；图中展示了调节分析所用的取值范围和关联强度。

图 C6

问题层面变量（探索性研究）



注。问题层面调节变量（难度：负荷-容量比；启发式最优性）的分布及相关性（皮尔逊相关）。展示这些信息是为了提高透明度，说明调节检验中的取值范围和潜在混杂。

附录D

跨研究汇总（森林图）

图D1

复杂度差异对选择的影响（*HC*、*CC*、*VC*、*DD*）的跨研究优势比（*OR*）及*95%* 置信区间。各点表示分别拟合于探索性与验证性数据的模型；虚线标示优势比 $= 1$。

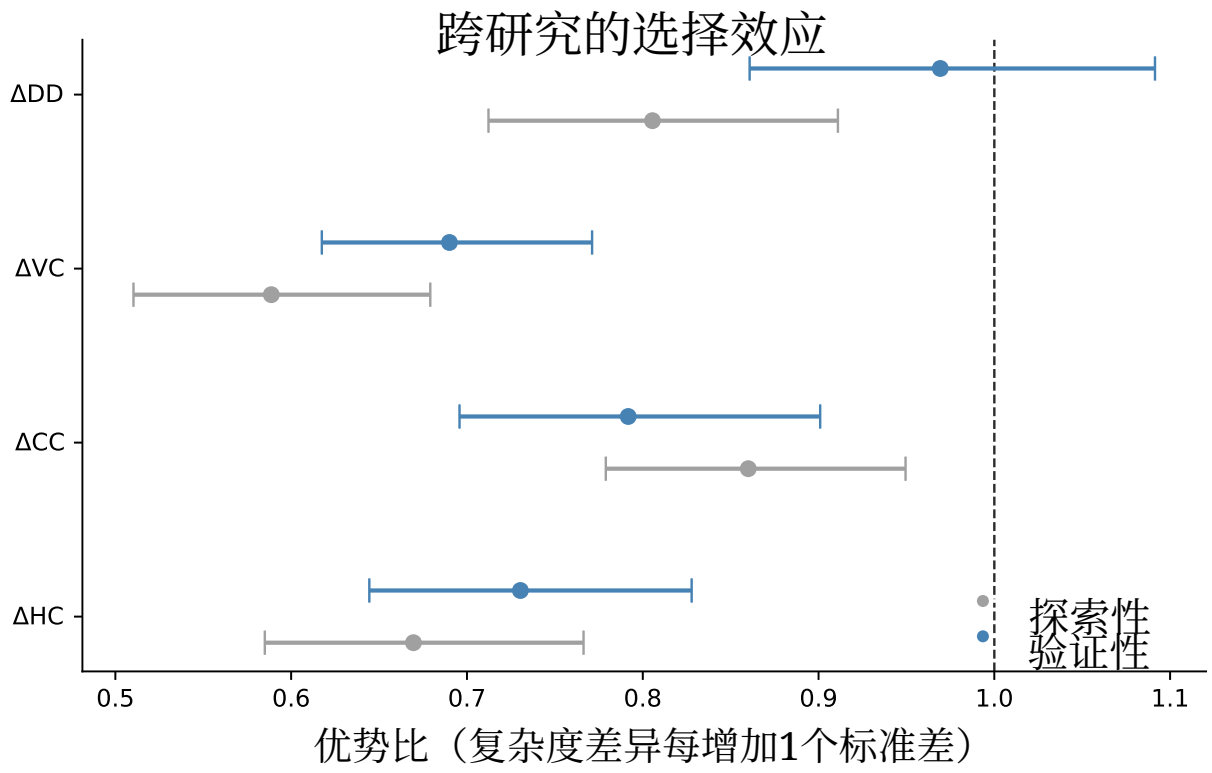


图 D2

在探索性与验证性研究中，*HC*、*CC*、*VC*、*DD* 每 *1* 个标准差绝对差值（|Δ|）所对应的反应时间百分比变化及*95%* 置信区间。虚线标示 *0%*。

