arXiv:2302.00707v1 [cs.HC] 1 Feb 2023

为什么结合文本与可视化能够改善贝叶斯推理：认知负荷视角

Melanie Bancilhon 圣路易斯华盛顿大学 圣路易斯，美国
mbancilhon@wustl.edu

AJ Wright
圣路易斯华盛顿大学
圣路易斯, 美国
ajwright@wustl.edu

Sunwoo Ha 圣路易斯华盛顿大学 圣路易斯，美国
sha@wustl.edu

Jordan Crouser 史密斯学院 北安普顿，美国
jcrouser@smith.edu

Alvitta Ottley
圣路易斯华盛顿大学
圣路易斯，美国
alvitta@wustl.edu

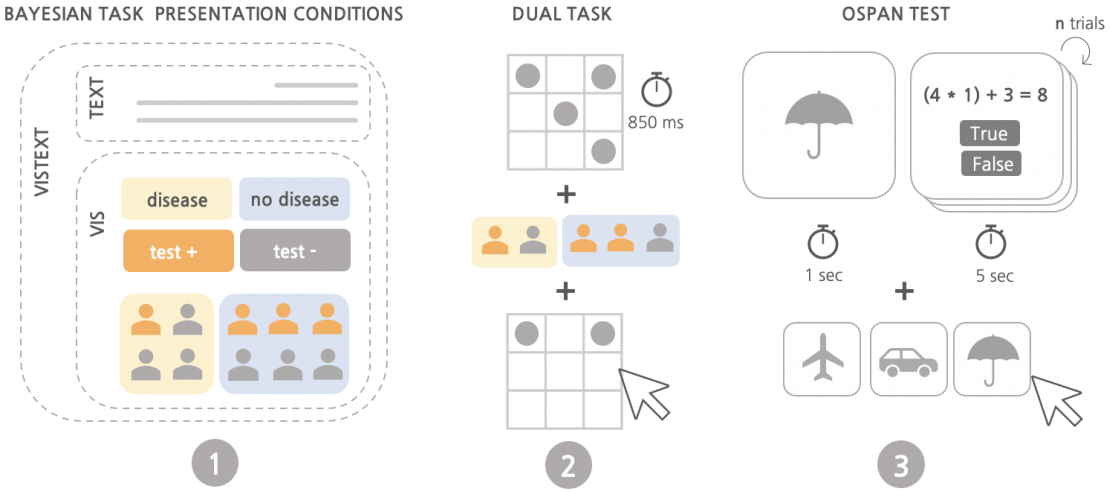

图1：我们的实验设计示意图。我们运用认知负荷理论研究呈现形式对贝叶斯推理的影响。1）向用户展示贝叶斯问题，形式包括：纯文本（文本）、纯可视化（可视化），或文本与可视化相结合（可视化文本）。2）我们采用双任务框架，通过要求参与者在完成贝叶斯任务的同时记住一个由4个点组成的图案，来操控认知资源。3）我们使用操作广度（ospan）任务测量工作记忆容量：该任务由 [18], 设计，通过测试参与者在完成穿插其中的数学问题时记住由四、五、六张图像组成的序列的能力来进行测量。

摘要

关于利用可视化改善贝叶斯推理并促进风险沟通的研究得出了不一的结果，这表明认知能力可能会影响用户在不同呈现形式下的表现。我们的研究考察了使用图标阵列、文本以及文本与图标阵列并置的方式解答贝叶斯问题时所引发的认知负荷。我们采用三管齐下的方法，全面考察认知需求，并测量工作记忆容量、双任务范式下分散注意时的表现，以及自我报告的努力程度主观评分。我们发现，工作记忆容量较低的个体在问题包含图标时错误更少，主观工作负荷也更低与仅使用文本相比，图表显示可视化能提高准确率，同时降低认知需求。我们认为，这些发现将对易于理解的风险沟通产生重大影响，尤其是对工作记忆容量较低的个体。

CCS 概念

• **贝叶斯推理** → **可视化**；决策；• **认知负荷** →多媒体；评估。

关键词

决策、贝叶斯推理、感知与认知负荷

ACM 参考文献格式：
Melanie Bancilhon、AJ Wright、Sunwoo Ha、Jordan Crouser 和 Alvitta Ot- tley。2023。为什么结合文本与可视化可能改善贝叶斯

CHI '23，2023年4月23日至28日，汉堡2023。ACM
ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

推理：认知负荷视角。载于*reCHInnect- ing* 论文集（*CHI '23*）。ACM，纽约州纽约，15 页。https://doi.org/10.1145/ nnnnnnn.nnnnnnn

1 引言

长期以来，学者们一直在研究多媒体形式对不同情境下理解和表现的影响。例如，在心理学领域，研究表明，将图表与文本描述结合起来，比单独展示其中一种更有助于学习（例如，[13, 14]）。同样，在教育领域，学者们提倡采用多媒体表征，而非单一形式 [37]。然而，对于可视化而言，相关指导原则并不那么明确，尽管在大众媒体叙事、教育和健康传播中，文本与可视化的结合十分普遍。

可视化研究中，文本与可视化结合的效果尚不确定的一个领域是条件概率的传播。条件概率或贝叶斯推理对于向广大受众传达重要统计信息至关重要，在医疗决策中尤其如此。具体而言，卫生官员需要说明检测结果显示某人感染病毒、但实际上并未感染的频率（假阳性）。此外，患者需要了解检测结果呈阳性时自己患病的概率（真阳性），以便就风险和潜在治疗方案做出明智的决定。尽管如此，大量研究表明，即便采用多媒体表征，无论新手还是专家，理解条件概率都颇具挑战性 [26, 32, 38, 73, 75]。

提升贝叶斯推理准确率最重要的指导原则之一，是以自然频数格式呈现信息（例如，*10* 个中有 *8* 个），而不是使用百分比（例如，*80%*）[32, 45, 56]。然而，进一步探究加入可视化是否能够改善贝叶斯推理的研究得出了不一的结果。早期研究发现，在文本格式中加入图标阵列等可视化内容，能够促使人们比阅读纯文本格式时更快、更准确地作答（例如，[11]）。较近开展的众包研究发现，只有在从文字描述中删除数值数据时，用欧拉图补充文本信息才会提高准确率 [49],，这表明同时呈现数字与可视化可能存在潜在冲突。研究人员考察了将文本与可视化关联起来的交互技术，但发现与静态多媒体形式相比，这些技术并未带来可衡量的益处 [51]。其他研究表明，空间能力是影响准确率的中介因素，并主张在可视化评估中考虑个体差异 [56]。

关于贝叶斯推理呈现形式的研究不仅限于可视化领域，其范围也远超我们在此重点介绍的少数几篇论文。然而，尽管相关研究十分丰富，我们的认识仍然有限，部分原因在于过度依赖推理准确性等粗略的绩效测量指标。我们认为，其他因素（例如不同呈现形式所引发的认知负荷）或许能为我们进一步揭示人们如何运用文本和可视化来支持贝叶斯推理的内在机制提供新的视角。认知负荷衡量的是特定任务对用户认知系统的影响 [57]。在各种情境下，它都会影响用户体验，例如在压力或情绪负担之下 [50], 分散注意时 [15, 16, 67], 或心理资源有限时 [2, 33]。

我们采用三种不同的方法评估图标阵列（仅可视化）、文本（纯文本）以及图标阵列与文本的组合（组合）所引发的认知负荷：工作记忆容量测试、双任务和自我报告的努力程度。我们认为，测量工作记忆容量有助于深入了解用户认知能力的个体差异。此外，双任务范式通过占用认知资源，能更直接地测量格式对认知负荷的影响，同时也模拟了注意力分散的现实情境。最后，我们通过 NASA-TLX 问卷收集了用户感知到的努力程度。这三种方法相结合，能够全面呈现认知负荷。

通过观察工作记忆容量的个体差异，我们发现，与纯文本格式相比，工作记忆容量较低的个体使用仅可视化 时错误显著更少。此外，NASA-TLX得分显示，与单独使用文本相比，工作记忆容量较低的用户表示，使用仅可视化 和组合 格式时，时间需求和体力需求较低。工作记忆容量较低的用户还表示，与纯文本相比，使用组合 时挫败感较低。综合来看，这些结果支持这样一种观点：与单独使用文本相比，可视化引发的认知负荷较低。

总而言之，本文记录了对可视化支持的贝叶斯推理研究的以下贡献：

（1）从认知负荷的角度出发，我们的发现为可视化在贝叶斯推理中的作用提供了新视角。特别是，我们发现**在组合格式中，以文本和可视化呈现重复信息可能会带来益处**。我们提供了初步证据，表明这**使人们能够选择更契合其心理模型的格式**。（2）我们证明，**工作记忆容量的个体差异会影响人们使用不同格式进行贝叶斯推理**。这对可视化在广泛总体中的应用（例如医疗决策）具有启示意义，也为可视化推荐过程增添了新的复杂维度。（3）我们展示了**如何使用不同的认知负荷测量方法进行可视化评估**，并通过扩展传统的准确率等性能指标，丰富了倡导多样化评估指标的相关研究。

2 背景

众所周知，人们不擅长根据条件概率进行推理 [3, 38]。例如，请看 [32] 中的以下情境：

> 一位参与常规筛查的*40*岁女性患乳腺癌的总体概率为*1%*。如果这位女性患有乳腺癌，她的乳腺*X*线摄影结果呈阳性的概率为*80%*。如果一位女性没有患乳腺癌，她的乳腺*X*线摄影结果也呈阳性的概率为*9.5%*。

乳腺*X*线摄影结果呈阳性。一位属于这一年龄组的女性在常规筛查中乳腺*X*线摄影结果呈阳性。

What is the probability that she actually has breast cancer?

根据贝叶斯定理，

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)} \quad (1)$$

在我们的情境中，D 是乳腺X线摄影结果呈阳性，H 是相关女性患有乳腺癌这一假设。人们（包括专家）常常会受到基率忽视的影响，在推理真阳性率 [38]时忽略基率 P（H）。几十年来，各领域一直致力于探索各种方法，通过减轻基率谬误来改善贝叶斯推理。

多项研究表明，频数格式（例如，10例中有8例，而非80%）有助于贝叶斯推理，并能显著提高准确率 [10, 23, 31, 32, 45]。此外，许多研究人员探讨了可视化对贝叶斯推理的影响（例如，[20, 21, 41, 42, 69]），其中最常见的设计是欧拉图 [11, 39, 41, 49] [8, 以30, 39, 41, 49, 55, 68, 71] 及频数网格或图标阵列。这些设计体现了贝叶斯促进背后的两种主导理论。欧拉图与嵌套集合理论相一致，有助于观察者推理子集之间的关系[5, 45, 68, 72], ；而展示自然频数（即，10例中有8例）的图标阵列则与生态理性框架 相一致，该框架基于进化理论提出，人类更擅长对可计数的对象进行推理[23, 31]。我们的研究采用图标阵列，因为它广受欢迎，而且自然频数格式在贝叶斯推理方面的成效已有充分记录（例如，[10, 23, 32, 49, 56]）。

为探究可视化在贝叶斯推理中的潜在益处，研究人员通常会比较以文本格式呈现的贝叶斯问题与以结合可视化和文本的格式呈现的问题所得的回答。然而，这些研究得出了不一的结果。例如，Micallef 等人 [49] 发现，单独使用文本与结合文本和可视化，在准确率方面没有可测量的差异。尽管如此，他们的后续研究表明，从文本中去除数字后，贝叶斯准确率显著提高。Ottley 等人 [54, 56] 重复了第一项研究的结果，发现单独使用文本与组合格式之间的总体准确率没有可靠差异。不过，他们发现，空间能力较高的参与者单独使用可视化时，表现可靠地优于单独使用文本 [56]。在另一项研究中，Ottley 等人 [54] 使用眼动追踪，考察人们在贝叶斯推理问题中如何从纯文本、可视化和组合格式中提取信息。他们发现，用户借助可视化能够轻松识别信息，但从文本中提取信息则更加容易。此外，他们的分析发现，研究参与者看到组合呈现时，使用各格式的方式并无差异。最后，Mosca 等人 [51]研究了通过交互将文本与可视化关联起来的效果。他们发现，与静态格式相比，加入交互并未提高贝叶斯推理的准确率。

我们认为，关于视觉设计能否提升贝叶斯推理这一悬而未决的问题，可能源于对其底层认知机制缺乏了解。Lesage 等人 [45] 的研究表明，无论呈现形式如何，贝叶斯推理表现都取决于可用的心理资源。尽管可视化研究人员常致力于提升速度和准确性度量，但我们对可视化对认知负荷的影响知之甚少。此外，在对可视化进行推理时，速度和准确率并不总是与认知负荷相关 [36, 65]。因此，我们有必要了解不同呈现形式下支配贝叶斯推理的过程。本文通过考察工作记忆容量的个体差异、采用双任务范式并测量主观认知负荷，拓展了对贝叶斯沟通技术的评估。我们旨在更细致地理解呈现形式对贝叶斯促进的潜在影响，并提供更全面的可视化设计指南。

2.1 测量认知负荷

工作记忆由多个组成部分构成，能够在有限时间内存储有限数量的信息，是推理过程中不可或缺的资源 [24]。认知负荷通常定义为处理一项任务所需的工作记忆量，是一项重要的可用性因素，可反映信息处理的难易程度 [57]。测量认知负荷的方法有很多，包括自我报告测量（例如 NASA-TLX）、基于表现的测量（例如双任务范式、操作广度测试）以及生理测量（例如瞳孔测量、功能性近红外光谱） [12, 25, 36, 40, 46, 52, 57, 59]。一些研究者运用这些方法探究可视化设计对认知负荷的影响 [1, 9, 18, 61, 70, 78], ，有时还重新审视长期以来的既有观点。例如，Matthews 等人强调，应采用多种方法交叉考察工作负荷的影响 [48]。Borgo 等人挑战了有关图表装饰冗余的传统观念，并通过双任务范式表明，与其他可视化相比，视觉装饰并不会引发更高的认知负荷 [9]。Peck 等人使用功能性近红外光谱和 NASA-TLX 评估可视化界面，发现条形图与饼图引发的认知负荷并无差异，这与普遍看法相反 [61]。

虽然生理测量已被证明是测量认知负荷的有效方法，但其侵入性较强，不适合在现实生活中应用 [43]。其他测量方法更易采用，有助于开展纵向研究，也使我们能够调查不同总体。在本研究中，我们采用三种不同的方法，探究呈现格式对贝叶斯推理认知负荷的影响：通过运算任务观察工作记忆容量的个体差异、双任务范式，以及通过 NASA-TLX 问卷收集自我报告评分。

2.1.1 认知负荷的个体差异方法。个体差异会影响我们如何运用不同格式进行推理（有关可视化中的个体差异的全面综述，请参见 [47] ），且有充分证据表明，认知特质会影响统计推理 [45, 51, 56, 76]。一些研究者发现，当信息以自然频数的形式呈现时，工作记忆容量高的参与者表现显著优于工作记忆容量低的

参与者 [45, 76]。Castro 等人 [17] 表明，在对不确定性可视化进行推理时，不同的可视化设计会引发不同程度的认知负荷。

一项与工作记忆容量测量结果呈高度相关的测试是认知反思测试（CRT）[45]。CRT 测试衡量一个人克服启发式与偏差并触发分析性思维的能力 [45]。测量工作记忆容量个体差异的一种更直接的方法是采用操作广度任务（OSPAN） [18]。在典型的 OSPAN 任务中，参与者必须在依次解答简单数学方程的同时，尝试按照正确顺序记住呈现的单词。本文采用 Castro 等人改编的在线操作广度测试来测量工作记忆容量 [17] [1]。作为对这一方法的补充，我们采用双任务范式；根据 Lesage 等人 [45], 的说法，这种范式可用于推断认知资源在贝叶斯推理任务表现中的因果作用。

2.1.2 双任务范式。尽管双任务方法尚未在可视化研究中得到广泛应用，但它是评估任务对认知资源依赖程度的有效方法，并且几十年来一直被心理学用于评估工作负荷 [19, 60]。在双任务范式中，用户同时执行两项任务，即主要任务和次要任务。这会造成分散注意并增加认知负荷，与仅执行主要任务相比，会导致表现下降。这种下降通常称为双任务成本 [58], ，可用于推断任务所引发的认知负荷。

一些研究人员使用双任务范式研究格式对认知负荷的影响 [9, 16, 70], ，其中一个原因是，这种方法有助于模拟现实生活条件，因为在现实生活中注意往往是分散的 [15, 16, 67]。Castro 等人 [16] 采用双任务方法，研究显示尺寸和移动设备屏幕大小如何影响注意。在他们的研究中，参与者通过倾斜移动设备，在不同尺寸的显示屏上控制一个蓝色球的移动（主要任务），同时执行变化检测任务：在固定显示屏上，口头报告4个箭头中哪些箭头改变了方向（次要任务）。通过这种方法，他们发现，在分散注意的情况下，较大的显示屏在心理上要求更高。Tintarev 等人 [70]使用双任务范式研究规划的呈现方式对认知负荷的影响。参与者需要在记忆中保持一份词语列表的信息，同时回答有关某项规划的问题，随后按正确顺序回忆这份词语列表。作者发现，在不同的规划格式之间，表现没有可靠差异。

在我们的研究中，我们采用受 [45], 启发的双任务方法，要求参与者在执行主要任务的同时，记住网格上的四点图案，以量化不同呈现形式所引发的认知负荷差异。

3 研究目标

我们设计了两项互为补充的研究，以探究认知负荷是否有助于解释贝叶斯推理与可视化方面相互矛盾、有时甚至令人困惑的研究发现。

综合来看，这些发现表明，认知资源与贝叶斯促进之间可能存在关联——在一项本就具有认知挑战性的任务中加入可视化和交互，或许无法产生预期效果。我们尚未充分了解认知负荷如何影响不同格式下的贝叶斯推理。基于此，本研究着重考察贝叶斯推理情境下，仅可视化、纯文本以及文本与可视化相结合这几种格式所引发的认知负荷是否存在差异。可视化条件采用图标阵列，因为它广泛用于传达贝叶斯信息，尤其是在医疗风险情境中，因此具有较好的生态效度。

在考虑实验设计方案时，我们权衡了以下因素：(1) 控制框架效应和学习效应，(2) 尽量减少个体差异带来的噪声，以及 (3) 尽量缩短研究的整体时长。遗憾的是，没有任何一种实验设计能够实现完美平衡。因此，我们报告了两项受控用户实验的结果。第一项采用组间设计，设置 3 种（呈现形式） × 2（负荷条件），以避免被试内研究可能引入的学习效应。第二项采用混合设计，其中 3 种（呈现形式）为组间因素，而× 2（负荷条件）为被试内因素，从而更好地控制个体差异。两项实验共同讲述了一个连贯的故事，揭示认知负荷、贝叶斯推理与可视化之间的关系。

4 实验1：被试间研究设计

我们将每位参与者随机分配到三种呈现条件之一——图标阵列（可视化）、文本（文本），或图标阵列与文本的组合（可视化文本）——从而在被试间比较不同的呈现形式。我们还将每位参与者随机分配到**单任务**或**双任务**条件，从而也在被试间比较这两种任务。我们选择被试间研究设计，是为了确保所有条件下的贝叶斯问题保持一致。以往研究表明，不同的贝叶斯情境会导致不同的准确率 [49] [2]。

4.1 呈现形式与贝叶斯任务

我们复现了 Mosca 等人 [51] 的分组图标阵列设计；在他们测试的可视化呈现形式中，该设计的准确率最高。作者根据 Bertin[6]的指南设计了图标阵列：用背景颜色区分总体中患病与未患病的成员，用图标颜色区分总体中检测结果呈阳性与阴性的成员。我们的文本条件下的参与者看到的是相同数据的文本格式，而可视化文本条件下的参与者则同时看到文本格式和图标阵列，两者垂直堆叠呈现。

我们向参与者展示了某种疾病在总体中的患病率数据，以及以可视化、文本 或可视化文本形式呈现的检测结果。我们请他们估算：i）检测结果呈阳性的人数；ii）在这些人中，实际患有该疾病的人数。这种依次询问用户

[1]本研究使用的 OSPAN 测试链接（由 [17] 开发）： https://bit.ly/2QHErIv

[2]实验1调查问卷、数据和分析的链接：https://bit.ly/3BFwokx

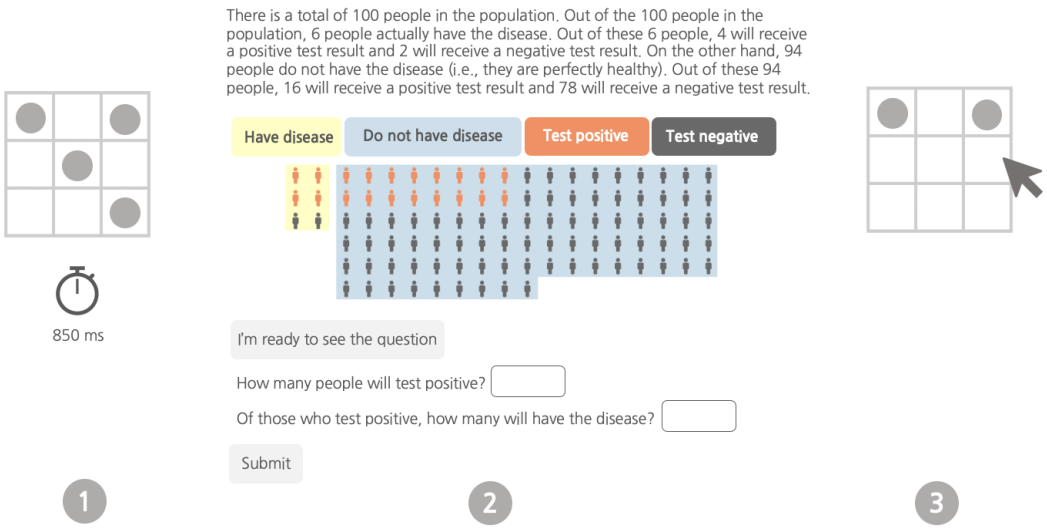

图2：贝叶斯调查概览：双任务条件下采用可视化文本格式。1）向用户展示由四个点组成的图案，该图案位于 3x3 网格上，展示时间为 850 毫秒，用户需记住该图案。2）这是可视化文本条件下贝叶斯任务的示例。用户需阅读问题，然后在准备好回答问题时按下按钮。3）用户提交贝叶斯问题的答案后，需在一个空白的 3x3 网格上重现点阵图案。

阳性人数和真阳性人数的技巧称为探测。探测 是一种有效的技巧，它通过提取嵌套数据（使用"其中"一词），独立于数学技能来评估贝叶斯理解。研究表明，与未采用探测的问题相比，这种技巧能引出更准确的回答 [11, 23, 56]。

4.2 负荷条件与双任务方法

参与者看到的要么是单独的贝叶斯概率估计任务，要么是该任务与一项次要任务的组合。被随机分配到双任务条件的参与者会看到一个由四个点组成的网格图案，持续 850 毫秒，并被要求在记住该图案的同时完成贝叶斯概率估计任务。随后，参与者需在空白网格上选择相应的单元格，尽可能准确地重现点阵图案。图1展示了双任务设置，其灵感源自 Lesage 等人关于 [45] 纯文本格式的研究，最初由 Bethell 等人开发。 [7]这项任务适用，因为它会占用视空工作记忆，可能干扰主要任务，从而造成预期的认知负荷增加。

4.3 能力测量与调查问卷

问卷还包含一份 NASA-TLX 问卷、一项空间能力测试和一项认知反思测试（CRT）。随后，参与者完成了来自 [18] 的一份工作记忆容量问卷。

NASA-TLX。 我们使用 NASA-TLX [34, 35] 来评估参与者的主观工作负荷。参与者报告了他们认为贝叶斯任务在以下六个子量表上的负荷：心理需求 时间需求、挫折感、体力需求、表现和努力程度。

工作记忆容量测试（OSPAN）。我们要求参与者依次记住一系列物体，同时回答简单的判断正误数学题。该测试包含 6 组序列，每组序列长度为 4 、 5 或 6 个，每种长度各呈现两次，顺序随机（术语 n 个跨度指序列出现n次）。在每个跨度中，参与者会看到一张图像，持续 1 秒，并被要求在 5 秒内回答一道简单的数学问题，同时将图像记在脑海中。该序列重复n次；每个跨度结束时，参与者必须按正确顺序回忆所呈现的图像。此版本的 OSPAN 由 Castro 等人设计 [16]。

认知反思测试。认知反思测试（CRT）已被证明是衡量认知负荷的有效方法 [45]。在我们的研究中，我们使用了一个包含 3 道题目的 CRT 版本，用于测试从类型1（直觉式）推理切换到类型2（策略式）推理的能力。由于后者需要运用工作记忆 [38], , 研究人员认为，能够完成这种切换的人具有较高的工作记忆容量 [59]。

空间能力测试。 空间能力测试用于衡量个体处理视觉和空间信息的能力。在本研究中，我们采用了 Ekstrom、French 和 Hardon [27] 编制的纸张折叠测试（VZ-2），分为两个环节，每个环节限时 3 分钟，共有 10 道题。其他研究已将该测试作为标准方法，用于评估不同空间能力水平下的贝叶斯推理表现 [39, 49]。

VisText 使用报告。 我们询问了可视化文本条件下的参与者，他们在回答贝叶斯问题时使用了多少比例的可视化和文本。他们报告了自己

偏好的方法，方法是从仅使用文本 到 仅使用可视化的刻度上选择相应的值

4.4 假设

H1 我们假设，贝叶斯推理任务中的表现取决于可用的认知资源。因此，**单任务** 条件下的推理准确性将高于**双任务** 条件。

H2 由于可用的认知资源受工作记忆容量影响，我们预计工作记忆容量高的个体比容量低的个体更准确，尤其是在**双任务** 条件下。

H3 先前研究考察了纯文本、图标阵列以及文本与图标阵列并置对贝叶斯推理的影响，发现三种呈现形式之间的准确率没有显著差异 [49, 56]。因此，我们预计在可视化、文本和可视化文本 的**单任务** 条件下，贝叶斯推理准确率没有显著差异。

关于预注册假设 H4a - H4d 的详细分析，请参阅补充材料。

4.5 参与者

我们通过亚马逊 Mechanical Turk 招募了来自美国、会说英语且 HIT 接受率为 100% 的用户。

报酬。所有参与者的报酬均符合最低工资法的规定，平均获得 4.84 美元，并花费 25.2 分钟完成两份问卷。在贝叶斯任务中，参与者每答对一道题可获得 0.50 美元的奖金。**双任务** 条件下的参与者还被分配了一项额外任务，因此需要花费更多时间完成任务，故每正确记住一个点可额外获得 0.25 美元（即比单任务条件下最多多获得 1 美元奖金）。按记住的点数发放奖金，也起到了激励参与者记住图案的作用。

我们使用 G*Power 软件针对混合方差分析进行了统计功效分析，确定要达到 95% 的统计功效，目标样本量为 251。由于 Mechanical Turk 研究通常有较高的排除率，我们招募了 450 名参与者。我们要求用户分别完成两份问卷：贝叶斯调查和 OSPAN 调查。根据先前研究确定的预注册排除标准 3 要求用户：i）每份问卷仅填写一次；ii）完成两份问卷；iii）在 OSPAN 测试的数学部分得分高于随机水平；iv）在 OSPAN 测试的记忆部分得分高于 10%；v）在点阵图案任务中的得分与平均值相差超过 2 个标准差。排除不符合排除标准的数据后，剩余 316 名参与者。初步分析数据后，我们发现了一些在预注册前未曾预料到的额外欺诈性和无效回答。我们决定排除在点阵图案回忆测验中输入超过 4 个点的用户，因为这会使他们更容易猜中正确图案（n=13）。我们还排除了对贝叶斯问题的回答小于或等于 0 的参与者（n=4），这表明缺乏

[3]实验1预注册链接：https://bit.ly/3xtC1zX

注意，并会在数据处理过程中导致无效错误值（见第 4.6 节）。我们进行了事后敏感性分析，结果表明，新增这两项排除标准并未影响研究结果（见补充材料）。在这些未预注册的排除之后，剩余 299 名参与者，其中 104 人被分配至文本组，100 人被分配至可视化组，95 人则看到可视化文本（其中 129 人处于**双任务**条件，170 人处于**单任务**条件）。

4.6 数据收集

本实验的自变量为：

- **3 呈现形式：** { 文本，可视化，可视化文本}
- **2 负荷条件：** { **单任务** ，**双任务** }

为了衡量贝叶斯表现，我们根据第4.1小节所述，计算了参与者响应的真阳性率。因变量包括：

- **准确** $\in \{0, 1\}$，用于表示响应是否准确的二元值。
- **偏差** 是响应与真实值的 log_{10} 比率。• **错误** 是偏差的绝对值。

虽然 exact 衡量的是逐字理解，但偏差和错误是对大意（近似）理解的替代指标，而推理和决策更常依赖这种理解 [4, 29, 39, 63, 64]。

协变量和其他计算得出的指标如下：

- **OSPAN** $\in [0...30]$，衡量一般认知能力。• **WMC** $\in \{low, high\}$，根据 ospan 得分的中位数二分法划分。• **NASA-TLX** $\in [0...20]$，衡量主观工作负荷的综合程度。
- **空间得分** $\in [-4...20]$，即空间能力测试得分。
- **空间水平** $\in \{low, high\}$，根据空间得分的中位数二分法划分。
- **CRT** $\in \{0, 1, 2, 3\}$，即认知反思测试得分。• **文本-可视化使用程度** $\in [1...20]$，对于处于可视化文本 条件的参与者，该指标取值范围为 0（主要使用文本）到 20（主要使用可视化）。

4.7 流失分析

关于在线研究中高流失率问题的研究日益增多 [44, 62, 77]。根据Zhou等人 [77], 的研究，对于认知负荷较高的研究，如果退出率达到或超过20%，就应当对此保持关注。作者还强调，应确保

表1：实验1中贝叶斯任务各条件下的退出率

负荷条件	退出率
单任务： 参与者完成了贝叶斯 Task	3.98%
双任务：参与者进行了贝叶斯任务和次要回忆任务	7.18%

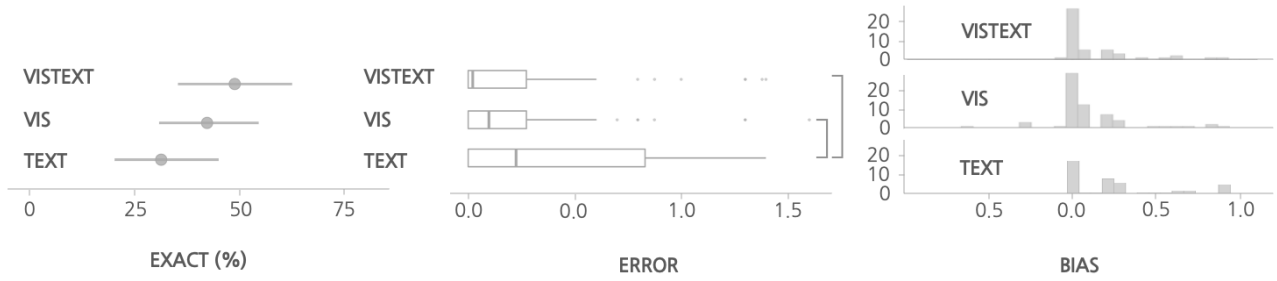

图3：单任务准确率（95%置信区间）、错误和偏差在不同呈现形式下的表现。] 表示两种形式之间存在显著差异（$\alpha = 0.0167$）。**我们发现错误在可视化和文本之间，以及可视化文本和文本之间存在显著差异。**

不同实验条件下的退出率没有显著差异，以此检查是否存在选择性流失。为提高透明度并鼓励有助于提升内部效度的做法，我们按照Zhou等人的建议进行了流失率分析 [77]。

我们的实验包含两份问卷：一项由作者设计实施的贝叶斯任务，以及一项来自 [18]、在 Qualtrics 上进行的 OSPAN 测试。我们对贝叶斯任务进行了流失分析，参与者被分配到单任务（单）或双任务（双）。我们借鉴了 Zhou 等人 [77]的方法，并且仅纳入同意参与研究的参与者；对于欺诈性回答，即参与者使用其记录的 IP 地址显示重复填写问卷的情况，则予以剔除。表1列出了各条件下的退出率，其计算方法依据[77]：将被分配到某一条件并完成全部任务的参与者人数 [4] 除以被分配到同一条件、至少表示同意且仅完成一次任务的参与者人数。我们观察到两种条件下的退出率都较低，且没有显著差异（χ^2（2） $= 1.88$， $p = 0.1704$， $d = 0.1406$）。

4.8 研究结果

在 299 名参与者中，104 人被分配到文本组，100 人被分配到可视化组，95 人则接受了可视化文本。每位参与者都完成了一个描绘第4.1小节中疾病 情境的单个贝叶斯问题。此外，170 人被分配到单任务（单任务 ）条件，129 人被分配到附加负荷的双任务条件（双任务 ）。

4.8.1 单任务：建立基线。W
e begin our analysis by examining
参与者在单任务（单任务）条件下完成任务，并检验呈现形式是否会影响表现。现有文献对于可视化对推理准确性的影响得出了不一的结果 [49,54, 56], ，而我们的**H3** 假设贝叶斯推理准确率不存在显著差异。

偏差。我们通过考察参与者的回答与准确答案之间的偏差程度，进行了探索性分析

并考察呈现形式对这种偏差的影响。我们观察到，单任务条件下的总体偏差中位数为 0.10；不同呈现形式的偏差中位数分别为：文本为 0.22，可视化为 0.00，可视化文本为 0.00。从图3可以看出，参与者的偏差并非正态分布。因此，我们在分析中使用非参数检验。此外，与使用文本的参与者相比，可视化和可视化文本条件下的参与者更有可能给出准确答案（偏差 $= 0$），但差异仅达到边缘显著。当我们以呈现形式作为被试间因素进行三组Kruskal-Wallis检验时，发现三种条件下的偏差存在显著差异（$H(2) = 12.87$， $p = .0016$， $\eta^2(H) = 0.065$）。采用校正后的显著性水平 $\alpha = 0.0167$ 进行后续Mann-Whitney-Wilcoxon检验后发现，可视化与文本之间（$W = 2379.5$， $p = 0.0006$， $\eta^2(H) = 0.092$）以及可视化文本与文本之间（$W = 1772.5$， $p = 0.0087$， $\eta^2(H) = 0.057$）的偏差均存在显著差异。

精确答案。我们考察了第一个准确率指标——精确答案，以探究呈现形式是否会影响正确答案的比例。总体而言，单任务条件下，40.9%的参与者正确回答了两个贝叶斯问题；在文本、可视化和可视化文本条件下，精确答案的比例分别为31.5%、43.08%和49.02%。总体比例 Z 检验表明，呈现形式对准确率没有显著影响（$\chi^2(2) = 3.4899$， $p = .1795$）。因此，精确推理成功的比例不受呈现形式影响。

错误 。 为了更细致地衡量准确率，我们考察了误差，以评估参与者的回答与精确答案相差多远，以及呈现形式是否会影响这一效果。总体误差中位数为 0.097；可视化条件下为 0.097，纯文本 条件下为 0.22，可视化文本 条件下为 0.021。Kruskal-Wallis非参数检验显示，各条件之间的误差存在显著差异（$H(2) = 8.43$， $p = 0.0148$， $\eta^2(H) = 0.037$）。采用校正后的显著性水平（$\alpha = .0167$）进行事后Mann-Whitney-Wilcoxon检验后发现，可视化与纯文本 （$W = 2227.5$， $p = .0093$， $\eta^2(H) = 0.0047$）之间，以及可视化文本 与纯文本 （$W = 1738.5$， $p = .0164$， $\eta^2(H) = 0.046$）之间存在显著差异。可视化与可视化文本之间没有显著差异（$W = 1610.5$， $p = 0.675$）。这些发现表明，使用纯文本进行推理所产生的错误显著多于其他格式。

总的来说，这些发现表明，呈现格式可能会影响推理错误。然而，观察到的影响

[4]当参与者完成整份问卷后，我们的服务器会记录实验结束时间戳

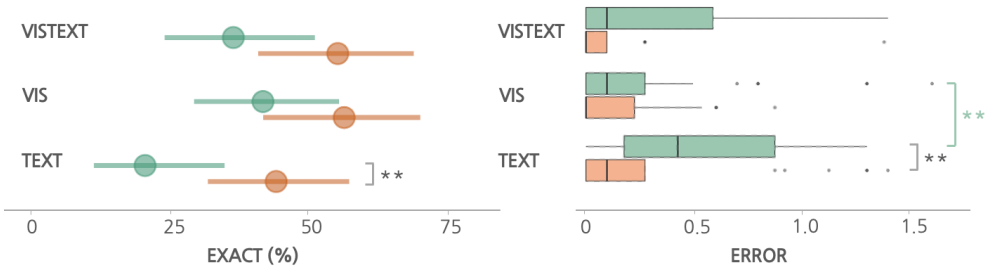

图 4：单任务任务准确率(95%置信区间)和错误在不同呈现形式和工作记忆组中的情况。 表示组间存在显著差异。我们发现，在该条件下，准确率和错误在低和高工作记忆组($\alpha = 0.05$)之间存在显著差异。在低工作记忆容量的个体中，错误在文本条件下显著高于可视化($\alpha = 0.0167$)

幅度较小，且对精确作答率没有显著影响。因此，**我们的结果仅部分支持 H3**。更具体地说，这些结果表明，与单独使用文本相比，即使将可视化与文本结合使用，也能提高贝叶斯准确率。值得注意的是，这些结果也与比较贝叶斯格式的可视化研究部分相悖。一方面，我们的发现与 Ottley 等人 [54, 56] 的研究相似：他们发现，文本、可视化和可视化文本在精确作答方面没有差异，但没有考察错误。另一方面，我们的结果与 Micallef 等人 [49]的研究不同：他们考察了文本和可视化文本，发现这些格式对精确作答或错误均没有可测量的影响。不过，我们的结果与先前研究之间的差异，可能归因于所用可视化类型以及实验设计的不同。

4.8.2 工作记忆容量中的个体差异。

本项目的主要目标是考察认知资源能否解释贝叶斯推理结果。具体而言，**H2**假设贝叶斯推理的准确率取决于可用的认知资源。为此，我们考察工作记忆容量对单任务中准确率的影响。

为了考察工作记忆是否会影响参与者准确率测量中的正确与错误情况，我们首先进行了二元逻辑回归，以检验 OSPAN 对正确作答的影响。结果发现，工作记忆测试每增加 5 分，贝叶斯问题答对的可能性就会提高至原来的 1.49 倍（95%置信区间 [.04, .12]）。对错误率的分析中，广义线性模型也显示 OSPAN 对错误率有显著影响（$t(169) = -3.326$, $p = 0.00108$）。因此，参与者的工作记忆容量越大，回答就越准确。

遵循以往研究 [18],我们根据参与者的 低 和高 工作记忆组是通过其中位数二分法划分的，依据是他们的 OSPAN 得分。图 4 总结了各工作记忆组在不同呈现形式下的准确率，并展示了各组相应的 完全 正确答案和 错误 分布比例。总体而言，在 单任务 中，高 组中有 52.29% 的参与者给出了 完全正确的 答案，而 低 组中这一比例为 34.59%。此外，低 组的中位数 错误 为 0.176，而 高 组的中位数 错误 为 0。与回归分析结果一致，我们发现 低 组和 高 组之间存在统计学上的显著差异，具体而言

我们比较了 完全正确答案 （χ^2（1） $= 6.4762, p = 0.0109, d = 0.3980$）和 错误 （Kruskal-Wallis检验， H（1） $= 6.8535, p = 0.0088, \eta^2$（$H$） $= 0.0348$）。这些结果共同支持**H2**，并显示 参与者的工作记忆可能会影响贝叶斯推理准确率。

4.8.3 工作记忆容量与呈现形式。
鉴于我们先前发现成功的推理可能取决于认知资源，我们进一步分析了呈现形式对低组和高组推理准确性的影响。具体而言，我们分别进行了三组比例检验，以比较答对频率，结果发现，低组（$\chi^2(2) = 4.8401, p = 0.0889$）和高组（$\chi^2(2) = 0.6504, p = 0.7224$）中，不同呈现形式之间均无差异。此外，在低组内对文本、可视化和可视化文本的错误进行比较的 Kruskal-Wallis检验显示，三种呈现形式之间存在统计学显著差异（$H(2) = 10.086, p = 0.006453, \eta^2(H) = 0.0817$）。我们使用校正后的显著性水平（$\alpha = 0.0167$）进行了成对的 Mann-Whitney-Wilcoxon检验，发现文本与可视化之间的错误存在显著差异（ $W = 874, p = .0017, \eta^2(H) = 0.1291$），但对于文本与可视化文本以及可视化与可视化文本之间的比较，未能拒绝零假设。在考察高组时，Kruskal-Wallis检验未发现呈现形式之间存在总体显著差异（$H(2) = 1.7301, p = 0.4210$）。这些分析表明，呈现形式的选择会影响工作记忆容量较低的用户；与可视化相比，文本会引发显著更高的错误率。然而，工作记忆容量较高的参与者受其所使用格式的影响较小。

我们在此进行的最后一项分析考察了低组和高组在各呈现条件下的表现。分析发现，在使用可视化进行推理时，低组和高组的精确回答比例（$\chi^2(1) = 1.2013, p = 0.2731$）及错误率（ $W = 428.5$ $p = 0.4381$ ，）相近。在使用可视化文本时，两个工作记忆组在精确回答比例（$\chi^2($ 1 $p =$ $) = 1.5875,$ ）0.2077 和错误率（ $W = 230$ $p = 0.1000$ ，）方面也没有差异。然而，我们观察到，在文本条件下，两组的精确回答比例（$\chi^2(1) = 5.889, p = 0.01524, d = 0.6997$）和错误率存在显著差异（$W = 235.5, p = 0.02481, \eta^2(H) = 0.0776$）。因此，与可视化或可视化文本相比，文本条件下低组和高组的准确率出现偏差的可能性略高。

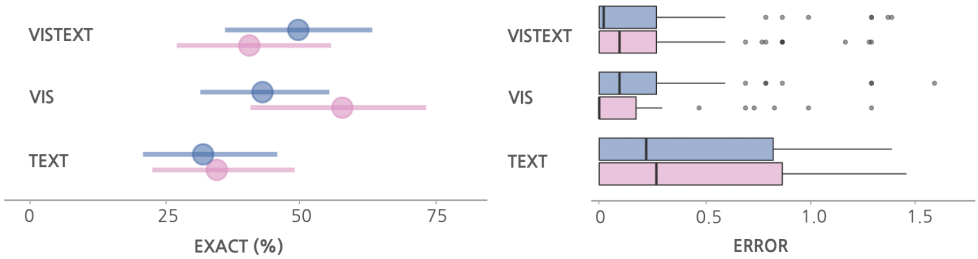

图5： 精确答案 (95%置信区间)和错误： 单任务 和 双任务 任务

4.8.4 双任务：分散注意下的推理。
在**H1**中，我们提出，如果通过增加一项次要任务来实验性地操纵执行能力，表现就会下降，这种现象称为**双任务成本**。因此，需要较高认知资源的格式会产生显著的双任务成本。

精确答案。我们观察到，单一条件与双重条件下的精确答案比例几乎相同。双重条件下的参与者有41.86%的时间给出了精确答案，而单一条件下这一比例为41.17%。我们比较了单任务与双任务中的精确答案比例，发现总体上没有显著差异（$\chi^2(2) = 0.0141, p = 0.9054$）。分析显示，文本的精确答案比例为34%，可视化为57.14%，可视化文本为38.64%。三样本比例检验发现，双重组中不同呈现形式的精确答案比例没有显著差异（$\chi^2(2) = 4.816, p = .09$）。因此，负荷操控对参与者的精确回答没有显著影响。

偏差。Kruskal-Wallis检验发现偏差在双任务组的不同呈现形式之间没有显著差异（$H(2) = 4.44$， $p = 0.1084$）。我们使用Mann-Whitney-Wilcoxon检验比较了偏差，发现单任务和双任务条件之间的总体差异不显著（$W = 11130$，$p = 0.8175$）。

错误。最后，我们还观察到，单任务和双任务的总体误差中位数均为0.097。使用Mann-Whitney-Wilcoxon检验进行总体比较后发现，单任务和双任务的错误没有显著差异（$W = 11232, p = 0.709$）。在双任务中，文本、可视化和可视化文本的误差中位数分别为0.27、0和0.097。与单一条件类似，总体Kruskal-Wallis检验显示，在双重条件下，呈现形式对错误有总体影响（$H(2) = 7.7344, p = .0207, \eta^2(H) = 0.0455$）。使用校正后的显著性水平（($\alpha = .0167$)进行后续Mann-Whitney-Wilcoxon检验，发现文本和可视化之间的错误存在显著差异（$W = 1162.5, p = .0073, \eta^2(H) = 0.0747$）。我们发现文本和可视化文本之间（$W = 1277.5$ $p = 0.1674$，$W = 612.5$ $p = 0.1002$）以及可视化和可视化文本之间（，）没有显著差异。

考虑工作记忆容量的差异。在第 4.8.2 节中，我们展示了工作记忆容量会影响贝叶斯推理的证据。在此，我们通过考虑工作记忆容量的个体差异，考察单任务和双任务条件下的表现差异。对于高组的个体，在考察单任务和双任务任务条件下的精确($\chi^2(2) = 0.4258, p = 0.5141$) 和错误($W = 3104, p = 0.3264$)时，我们未发现显著差异。同样，在考察单任务和双任务条件下低组参与者的精确($\chi^2(2) = 0.0701, p = 0.7912$) 和错误($W = 2467, p = 0.5253$)时，我们也未发现可测量的差异。综上，次要任务并未产生预期结果，且支持**H1** 的证据尚无定论。

尽管我们在第4.7节中发现，单任务与双任务的流失率没有显著差异，但我们进行了克鲁斯卡尔-沃利斯检验，以调查排除所有数据质量排除项后，两项任务中的OSPAN得分分布是否存在差异。我们发现，OSPAN得分在单任务与双任务($W = 13688$， $p = 0.0002$， $\eta^2(H) = 0.0422$)之间存在显著差异，OSPAN得分在双任务条件下更高。这可能是选择性流失或样本偏差所致。双任务条件下参与者的OSPAN得分显著高于单任务任务。这也可以解释为何我们没有观察到两项任务之间表现显著下降。在解读实验1的结果时，我们将考虑这一混杂因素。

4.8.5 ***NASA-TLX*** 自我报告的努力程度。
考察单任务任务中的自我报告努力程度时，我们发现，不同呈现格式下感知到的**挫折感**总体存在显著差异（Kruskal-Wallis检验， $H(2) = 11.72$, $p = 0.003$， $\eta^2(H) = 0.0582$）。我们针对成对比较进行了单独的Mann-Whitney-Wilcoxon检验，并采用校正后的显著性水平 $\alpha = 0.0167$；结果显示，**挫折感**在可视化文本与文本($W = 1918.5$, $p = 0.0005$， $\eta^2(H) = 0.1078$)之间存在显著差异。

由于工作记忆容量可能会影响自我报告的 NASA-TLX 得分，我们分别观察了各工作记忆组中不同呈现形式之间的差异。我们发现，在NASA-TLX的任何子量表上，高工作记忆组中不同呈现形式之间均无显著差异。在低工作记忆组中，我们分别进行了克鲁斯卡尔-沃利斯检验，发现以下指标的准确率在不同呈现形式之间存在显著差异：

- 时间需求：$(H(2) = 10.305, p = 0.006, \eta^2(H) = 0.0839)$
- 体力需求：$(H(2) = 8.95, p = 0.0114, \eta^2(H) = -0.0045)$
- 挫折感：$(H(2) = 10.825, p = 0.004, \eta^2(H) = 0.089)$

作为后续分析，我们采用校正后的显著性水平进行了Mann-Whitney-Wilcoxon检验 （$\alpha = 0.0167$），在低组中发现可视化文本与文本之间在以下方面存在显著差异：

- *temporal demand* ($W = 644, p = 0.004, \eta^2(H) = 0.1262$),

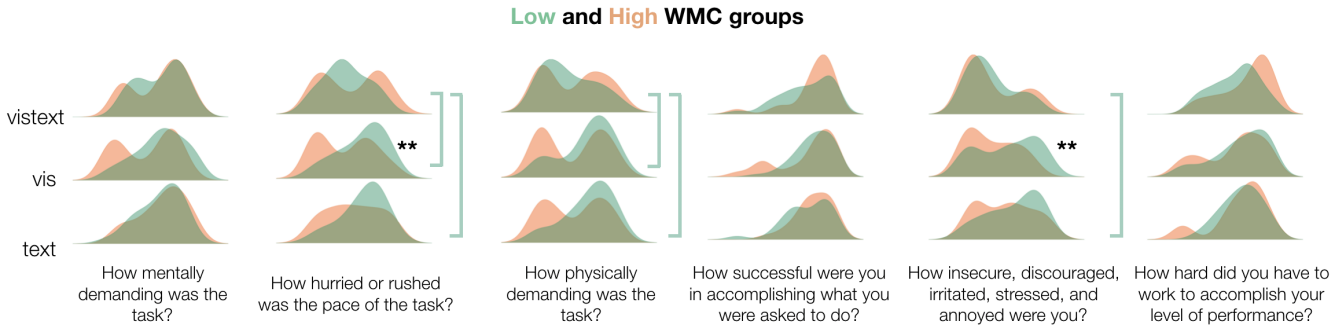

图 6： NASA-TLX 得分分布单任务 中低 和 高 工作记忆容量 组。 表示工作记忆组之间存在统计学显著差异 ($\alpha < 0.05$) 而] 表示相应工作记忆容量工作记忆容量 组($\alpha < 0.0167$)中不同格式之间的差异。**

- *physical demand* ($W = 638, p = 0.005, \eta^2(H) = 0.1175$)
- *frustration* ($W = 672.5, p = 0.0009, \eta^2(H) = 0.1712$)

在低组中，我们还发现可视化与可视化文本之间在以下方面存在差异：

• 时间需求 ($W = 894.5, p = 0.006, \eta^2(H) = 0.0904$), • 体力需求 ($W = 878, p = 0.011, \eta^2(H) = 0.08296$)

我们考察了每种呈现形式下NASA-TLX 得分在高 和 低 组之间的差异。在可视化条件下，我们发现以下差异：

• 时间需求 ($W = 307, p = 0.01525, \eta^2(H) = 0.0673$) • 挫折感 ($W = 332.5, p = 0.0379, \eta^2(H) = 0.0525$)

最后，我们发现，在文本 和可视化 条件下，工作记忆组之间报告的分数没有差异。

4.8.6 补充分析。

空间能力。 我们采用逻辑斯蒂链接函数拟合了广义线性模型，发现空间能力得分对精确性有显著影响（$z(298) = 4.670$， $p = 3.00e{-}06$）。我们还通过广义线性模型考察了空间能力得分对错误的影响，发现其效应显著（ t 298 $p =$ ( ) $= -4.003,$ $7.91e{-}05$ ）。这些发现重现了先前研究的结果，即空间能力会影响贝叶斯推理[51, 56]。

完成时间。 Kruskal-Wallis检验显示，呈现形式（$H(2) = 1.2454$， $p = 0.5365$）或负荷条件（$H(2) = 2.03$， $p = 0.1546$）对贝叶斯任务的完成时间均无显著影响。此外，我们发现，低 和 高 工作记忆组之间的完成时间没有显著差异（$H(2) = 0.0797$， $p = 0.7778$）。

认知反思测试。 我们的 CRT 结果在很大程度上重复了 OSPAN 的发现。我们发现，CRT 得分对精确答案有总体显著影响（$\chi^2(3) = 20.502$， $p = 0.0001$， $\eta^2(H) = 0.5246$）。此外，克鲁斯卡尔-沃利斯检验显示，CRT 对偏差（$H(3) = 20.566$， $p = 0.0001$， $\eta^2(H) = 0.0595$）和错误（$H(3) = 24.986$， $p = 1.555e{-}05$， $\eta^2(H) = 0.0745$）有统计学显著影响，这表明 那些 *CRT* 得分较高的个体更有可能输入精确答案，且推理错误更少。

5 实验2：混合设计研究 UDY

实验1采用被试间设计来控制学习效应，并通过比较参与者对同一贝叶斯问题的回答来确保一致性。然而，我们发现双任务对准确率没有显著影响。这可能是因为两组的工作记忆容量存在差异，也可能是研究设计导致个体差异较大。我们开展了第二项混合设计研究[5]，以 1）控制单任务和双任务中的个体差异；2）检验未能复现结果是否源于总体差异或方法差异。

为降低任务的整体难度并更好地控制个体差异，我们对实验设计作出了以下调整。

• **通过练习环节改进研究准备**：我们添加了一个研究前试次，让参与者熟悉任务和研究流程。在继续进行正式任务之前，参与者先观看并尝试完成一个贝叶斯推理任务样例。

• **控制个体差异**：我们采用混合因子设计，将负荷条件（单任务 、双任务 ）设为被试内因素，将呈现形式（可视化、文本、可视化文本）设为被试间因素。

• **移除CRT测试**：我们在实验1中发现，认知反思测试与 OSPAN 测试正相关（ r 297 $p =$ ( ) $= 0.27,$ $2.051e^{-06}$ ），而 OSPAN 测试的认可度更高，因此我们从问卷中移除了认知反思测试 [22, 28, 53, 74]。这缩短了问卷。

5.1 任务与流程

我们的任务与实验1相似，但在贝叶斯调查中，用户以不特定的顺序分别完成了单任务和双任务；贝叶斯问题则通过两个情境呈现：出租车情境和课程情境。解答贝叶斯

[5]实验2问卷、数据和分析链接：https://bit.ly/3LfgXCZ

表2：实验2中贝叶斯任务使用的情境

情境	描述
cab	共有100名目击者目睹了这起汽车事故。100名目击者中 , 15人声称造成事故的汽车是一辆出租车。在这15名目击者中 , 12人声称这辆汽车是蓝色的，3人声称这辆汽车是绿色的。在另一方面 , 85名目击者声称，肇事汽车不是出租车。在这85名目击者中，3人声称汽车是蓝色的，82人声称汽车是绿色的。
类别	总体中共有 100 名大学新生。在这 100 名新生中，30 名正在修读创业导论课程。在这 30 名新生中，20 名计划毕业后创业，另外 10 名则不打算创业。另一方面，有 70 名新生没有选修了创业导论课程。在这 70 名大一新生中，10 人计划毕业后创业，另有 60 人不打算创业 not.

问题后，用户完成了NASA-TLX和空间能力测试。在本实验中，我们没有进行认知反思测试。与实验1类似，用户还完成了OSPAN测试。

练习环节。在练习环节中，用户练习了点阵回忆任务、单任务 贝叶斯任务，以及将两项任务结合进行的双任务 条件。

报酬。 参与者获得 2 美元的基本报酬，并有机会赢得总额最高为 2.5 美元的奖金，其中每答对一道贝叶斯问题可得 0.5 美元，每正确复现一个点阵图案可得 0.5 美元。参与者平均获得 1.51 美元奖金，并平均用时 26.8 分钟完成贝叶斯问卷和 OSPAN 调查问卷。

5.2 实验设计

与实验1类似，我们将每位用户随机分配至三种呈现条件之一（可视化、文本或可视化文本），因此呈现方式采用被试间比较。每位用户都完成了单任务和双任务，并接触了出租车或汽车情境，因此负荷条件是被试内条件。

5.3 呈现条件

我们的呈现形式仍为组间条件，与实验1相同：文本、可视化和可视化文本。我们使用疾病情境进行任务前教程，每位参与者阅读两个描述不同情境的贝叶斯问题：出租车和课堂 [32, 49, 54]。出租车情境涉及目击者对肇事逃逸情境的证词，而课堂情境则呈现大学生的职业前景。我们将其中一个情境随机分配给单任务条件，另一个分配给双任务条件，并对条件顺序进行平衡。

5.4 参与者

根据我们的预注册6，我们基于三因素混合方差分析进行了功效分析，并确定理想样本量为168。为应对30%至40%的排除率，我们通过亚马逊 Mechanical Turk 招募了240名参与者。参与者均为来自美国的英语使用者，且HIT 接受率为100%。按照实验1中确定的相同预注册标准（见第4.5节）排除88名参与者后，最终保留152名参与者（文本= 46，可视化=55，可视化文本=51）。

5.4.1 流失率。我们采用与实验1相同的方法，对实验2进行了流失率分析。表3按条件列出了退出率，即首先看到双任务或单任务的参与者。我们发现两种条件之间的退出率没有显著差异($\chi^2(2) = 1.6824$，$p = 0.1946$，$d = 0.1293$)。排除部分数据后，我们考察了参与者在OSPAN测试中的表现，发现得分没有显著差异。这表明完成实验的总体在两种条件下保持一致（$H(2) = 0.34157$，$p = 0.5589$，$\eta^2(H) = -0.0048$）。

表3：实验2贝叶斯任务各条件下的退出率

任务顺序	退出率
单任务 、双任务 ：参与者先完成双任务随后完成单任务	11.23%
双任务、单任务：参与者先进行单任务随后进行双任务	15.67%

5.5 结果

在本实验中，我们旨在通过一项混合设计研究，揭示不同呈现形式所引发的双任务成本差异。首先，我们确定单任务准确率的基线，并将结果与实验1进行比较。然后，我们考察并比较不同呈现形式下由双任务（**双任务成本**）引发的表现下降。

5.5.1 单任务。

精确答案。总体而言，在单任务中，38.2%的参与者正确回答了两个贝叶斯问题；文本、可视化和可视化文本的精确答案比例分别为21.7%、47.3%和42.3%。与实验1相反，我们的分析显示，呈现形式对精确答案有显著影响（$\chi^2(2) = 7.58, p = .023, d = 0.4584$）。采用调整后的显著性水平（$\alpha = 0.0167$）进行后续两样本比例检验后发现，可视化与文本在精确答案方面存在显著差异（$\chi^2(2) = 7.1195, p = 0.0076, d = 0.5537$）。

偏差 。 我们发现，不同呈现形式之间的偏差没有显著差异（Kruskal-Wallis检验，H 2 p =( ) = 1.0826,, η^2( ) ）。0.582 H = −0.0063

[6]实验2预注册链接：https://bit.ly/3qIzJJu

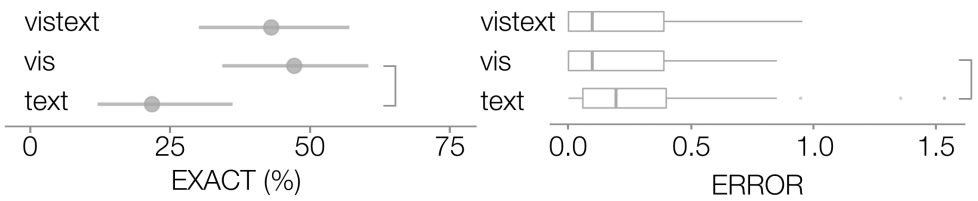

图 7：实验2。单任务任务精确 和错误在不同呈现形式下。] 表示两种格式之间存在显著差异（$\alpha = 0.0167$）。

错误 。 总体误差中位数为0.097；可视化条件下为0.097，文本条件 下为0.194，可视化文本条件 下为0.076。Kruskal-Wallis非参数检验发现，不同呈现形式之间的错误存在显著差异（ H 2 p = ( ) 0.0223 H = 0.0376 = 7.61,, η^2( ) ）。采用校正后的显著性水平（$\alpha = 0.0167$）进行事后Mann-Whitney-Wilcoxon检验后发现，可视化 与文本 之间存在显著差异（$W = 1619.5, p = 0.01329, \eta^2(H) = 0.05182$）。总体而言，整体趋势与实验1一致，并表明与可视化相比，参与者使用文本时的准确度最低。不过，实验2中的差异更为明显。

先前研究表明，不同的贝叶斯情境可能对准确率产生不同影响 [49]。我们发现，在准确率方面，出租车和出租车情境之间没有显著差异；具体而言，在精确度（$\chi^2(2) = 0.0251$，$p = 0.8741$，$d = 0.0257$）、偏差（$W = 2870.5$, $p = 0.9727$，$\eta^2(H) = -0.0067$）或错误率（ W = 2952.5 p = 0.7844 η^2 H = −0.00616 , , ( ) ）方面均如此。

5.5.2 单任务与双任务。

双任务。在双任务中，我们发现回忆出的点的平均数量为3.56（$\sigma = 0.77$）。40.8% 的参与者正确回答了两个贝叶斯问题；文本格式的精确答案比例为 26.1%，可视化格式为 41.8%，可视化文本格式为 51.9%。不同呈现形式之间的精确答案总体差异显著（$\chi^2(2) = 6.8197, p = 0.0331, d = 0.4335$）。后续两两比较显示，文本与可视化文本之间的精确答案存在显著差异（$\chi^2(2) = 6.8003$, $p = 0.0091, d = 0.5461$），这表明在分散注意的情况下，与单独使用文本相比，结合可视化与文本可以减少错误。我们发现，不同呈现形式在偏差（Kruskal-Wallis检验， $H(2) = 2.3795$, $p = 0.3043$）或错误（Kruskal-Wallis检验， $H(2) = 4.451$, $p = 0.108$）方面均无显著差异。

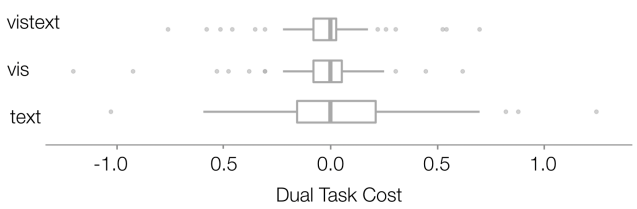

图 8：不同呈现形式下的双任务成本。

双任务成本。我们通过计算错误的差值，观察每位参与者在双任务中的表现相较于单任务的下降程度；这一指标称为**双任务成本**。通过比较不同呈现形式下的双任务成本，我们可以推断认知负荷的差异。我们进行了Kruskal-Wallis检验，发现不同呈现形式之间的双任务成本总体上没有差异（$H(2) = 1.0314$, $p = 0.5971, \eta^2(H)$ = −0.0065 ）。

校准双任务成本。在第 4.8.2 节中，我们展示了工作记忆容量会影响贝叶斯推理的证据。为此，我们观察了工作记忆容量高和低的组别中双任务成本的影响。我们发现，在高组中，不同呈现形式之间的双任务成本没有显著差异（Kruskal-Wallis 检验，$H(2) = 1.4324$，$p = .4886$，$\eta^2(H) = -0.0089$），在低组中也如此（Kruskal-Wallis 检验， $H(2) = .14215$, $p = .9314$, $\eta^2(H) = -0.0226$）。因此，我们得出结论：即使考虑工作记忆容量的个体差异，双任务的影响在不同呈现形式下也保持一致。

6 讨论

我们的研究借助认知理论，探究了贝叶斯传播情境下关于结合文本与可视化效果的相互矛盾发现。仅分析准确率的总体趋势，往往无法呈现评估研究的全貌，因为已有大量研究探讨个体差异对贝叶斯推理及其他方面的影响 [18, 51, 56, 76]。我们的结果表明，结合可视化与文本并不会增加认知负荷，在某些情况下还能改善主观工作负荷。以下介绍这些研究的主要结论。

我们分析了准确率，以便将研究结果与先前工作进行比较，并为认知负荷结果提供背景。从总体上看，我们得到的结果与可视化领域先前研究的结果相似 [49, 51, 54, 56]。 单凭呈现形式对贝叶斯推理几乎没有影响，但加入可视化后，贝叶斯推理能力有所提升。在实验1中，我们发现基线单任务条件下，用户答对的比例在三种格式之间没有显著差异，这复现了Ott-ley 等人的研究结果 [56]。与纯文本相比，用户使用仅可视化时的错误率显著降低，但该统计检验的效应量较小。在实验2中，与仅使用文本相比，采用组合呈现形式时答对比例显著更高，不过效应较小。

尽管准确率分析能够很好地揭示差异，但它并不能解释这些现象。

6.1 认知负荷的启示

我们采用了三种不同但互为补充的方法来评估认知负荷：工作记忆容量测试、自我报告的努力程度以及双任务。我们对工作记忆容量的研究受到既有研究的启发，这些研究主要关注纯文本格式，并表明工作记忆容量与推理表现呈正相关 [45, 76]。我们的研究发现，工作记忆容量的影响依然存在：工作记忆容量高的个体通常比工作记忆容量低的个体表现更好。这种影响在纯文本条件下尤为明显。这些发现有助于理解我们关于贝叶斯任务准确率的研究结果；结果表明，可视化和多媒体形式可能优于纯文本形式。

我们还发现，工作记忆容量低的参与者单独使用可视化时，比单独使用文本时表现更好。与 Castro 等人的研究一致 [17], , 工作记忆组中这种表现差异间接表明，纯文本引发的认知负荷高于仅可视化。进一步推论，文本与可视化的组合并未引发高于仅可视化的认知负荷。鉴于先前研究发现，在组合呈现中删除文本里的数字会对推理表现产生积极影响[49], , 我们原以为会找到证据，表明结合文本与可视化会增加认知负荷，但我们的数据并不支持这一观点。这些发现对可视化推荐和无障碍性具有实际意义。可视化有助于认知能力较低的群体，也适用于认知负担较重的情境。

值得注意的是，在实验1中，与仅使用文本相比，工作记忆容量较低的参与者在使用组合形式时报告的挫折感和时间需求显著较低。这一发现进一步支持了多媒体形式的使用。然而，值得注意的是，实验1还显示，对于工作记忆容量较低的个体，组合形式与文本在准确率方面没有显著差异，这凸显了仅分析准确率的不足。

总体而言，我们的发现一定程度上支持多媒体效应这一观点。

具体而言，同时提供文本和可视化可能有助于促进推理，尤其是对工作记忆容量较低的人而言。一种可能的解释是，可视化使观看者能够将信息从记忆中卸载，而文本则熟悉且易于处理。这一假设与先前在参与者解决贝叶斯任务时采集眼动数据的研究结果一致 [54]。这些研究结果表明，可视化有助于识别相关信息，但与视觉形式相比，文本可能更易于处理 [54]

另一个对我们结果的合理解释是，工作记忆较低的参与者可能更喜欢 com 的灵活性 合并格式使他们能够选择最符合其心理模型或偏好的格式。要更深入地理解这一现象，还需要进一步研究。

6.2 关于双任务范式为何失效

双任务范式未能揭示不同格式之间的认知负荷差异，即使将工作记忆容量的个体差异纳入考量也是如此。具体而言，要求参与者在记忆中保持一个点阵图案，并未影响他们的推理表现。我们在实验1中假设，这种效应可能是被试间设计中的个体差异所致。然而，被试内实验2得出了类似发现，这与H3及既往研究相矛盾 [45]，这可能是由于实验设计存在差异。例如，Lesage 等人[45] 开展了一项实验室实验，179名心理学一年级学生参与了此前的研究，并获得课程学分。我们的研究采用了人口统计构成更多样化的众包研究人群。另一种可能的解释是，次要任务过于简单，或者我们的研究参与者可能把图案写了下来，而不是在记忆中保持它。此外，观察到的差异也可能源于研究总体的人口统计构成不同。

一些研究人员制定了选择合适次要任务的指南，其中包括对任务难度和相似性的考量 [58, 66]。然而，要在主要任务和次要任务之间取得恰到好处的平衡颇具挑战，因为次要任务必须足够困难，才能增加认知负荷，但又不能难到导致认知过载。尽管复现失败的确切原因尚不清楚，我们仍鼓励研究人员在开展众包评估时考虑调整双任务方法。一种可选的研究设计是根据参与者的能力校准次要任务的难度。例如，Castro 等人 [18] 在一项探究分散注意对驾驶影响的研究中采用了校准。参与者先进行预试，以确定能使次要任务准确率达到 75% 的难度级别。对于我们选择的次要任务，可以根据参与者的表现，校准需要记忆的点阵图案的大小。或者，Borgo 等人关于视觉修饰对参与度和工作记忆影响的研究采用了词语选择次要任务 [9], , 要求用户从滚动词语列表中识别水果。研究人员可以通过为每位参与者调整词语的滚动速度来校准次要任务。此外，Borgo 等人 [9] 的双任务设置较不容易受到研究任务违规的影响，因为其中不涉及回忆任务。

7 结论

我们的工作通过考察和比较三种呈现形式，拓展了对工作记忆容量与贝叶斯推理之间关系的理解。通过考察更细致的准确性度量，我们表明，与纯文本格式相比，仅可视化和组合形式在贝叶斯推理中的错误更少。此外，我们表明，工作记忆容量会调节贝叶斯推理准确率，尤其是在文本呈现形式中。最后，我们表明，工作记忆容量较低的用户在文本呈现形式中的贝叶斯推理准确率尤其受到影响。

工作记忆容量较低的用户单独使用可视化时，比单独使用文本时得出的结论更准确。我们探讨这些发现如何影响可视化设计指南，尤其是针对工作记忆容量较低的用户。为此，我们主张采用更多样化的评估指标，并鼓励可视化领域

借鉴并应用认知科学及相关领域的现有研究，更好地理解人们如何感知可视化并基于可视化进行推理。

致谢

作者谨向 Lace Padilla 致谢，感谢她对认知方法应用提出的见解；也感谢 Shayan Monadjemi 对稿件提供的宝贵反馈。本材料所述工作得到了美国国家科学基金会的资助，资助编号为 2142977。

参考文献

[1] Erik W Anderson、Kristin C Potter、Laura E Matzen、Jason F Shepherd、Gilbert A Preston 和 Cláudio T Silva。2011。使用脑电图和认知负荷研究可视化效果的用户研究。载于*Computer graphics forum*，第 30 卷。Wiley Online Library，791–800。[2] Amy L Atkinson、Alan D Baddeley 和 Richard J Allen。2018。记住一些还是全部？衰老与视觉工作记忆中的策略效应。*Quarterly Journal of Experimental Psychology* 71，7（2018），1561–1573。[3] Laurel C Austin。2019。医师与非医师对诊断乳腺X线摄影与大规模筛查乳腺摄影中阳性预测值的估计：对贝叶斯推理的考察。*Medical Decision Making* 39，2（2019），108–118。[4] Melanie Bancilhon 和 Alvitta Ottley。2020。你领会要点了吗？理解可视化如何影响决策。*arXiv preprint arXiv:2010.04096*（2020）。[5] Aron K Barbey 和 Steven A Sloman。2007。基率尊重：从生态理性到双重过程。*Behavioral and Brain Sciences* 30，3（2007），241–254。[6] Jacques Bertin。1983。图形符号学：图表、网络、地图。技术报告。[7] Charles E Bethell-Fox 和 Roger N Shepard。1988。心理旋转：刺激复杂性与熟悉度的影响。*Journal of Experimental Psychology: Human Perception and Performance* 14，1（1988），12。[8] Katharina Böcherer-Linder 和 Andreas Eichler。2019。如何提升贝叶斯推断任务中的表现：五种可视化的比较。*Frontiers in psychology* 10（2019），267。[9] Rita Borgo、Alfie Abdul-Rahman、Farhan Mohamed、Philip W Grant、Irene Reppa、Luciano Floridi 和 Min Chen。2012。一项关于在可视化中使用视觉修饰的实证研究。*IEEE Transactions on Visualization and Computer Graphics* 18，12（2012），2759–2768。[10] Gary L Brase。2008。对模糊统计信息的频率解释有助于贝叶斯推理。*Psychonomic Bulletin & Review* 15，2（2008），284–289。[11] Gary L Brase。2009。统计推理中的图示表征。*AppliedCognitive Psychology: The Official Journalof the Society for Applied Researchin Memory and Cognition* 23，3（2009），369–381。[12] Roland Brunken、Jan L Plass 和 Detlev Leutner。2003。多媒体学习中认知负荷的直接测量。*Educational psychologist* 38，1（2003），53–61。[13] Tad T Brunyé、Holly A Taylor 和 David N Rapp。2008。程序性多媒体演示中的重复与双重编码。*AppliedCognitivePsychology:The Official Journal of the Society for Applied Research in Memory and Cognition* 22，7（2008），877–895。[14] Tad T Brunyé、Holly A Taylor、David N Rapp 和 Alexander B Spiro。2006。学习程序：工作记忆在多媒体学习体验中的作用。*Applied Cognitive Psychology* 20，7（2006），917–940。[15] Alan D Castel、Jay Pratt 和 Fergus IM Craik。2003。空间工作记忆在返回抑制中的作用：来自分散注意任务的证据。*Perception & Psychophysics* 65，6（2003），970–981。[16] Spencer Castro。2017。手持移动设备的尺寸和手部位置可能如何影响分散注意。载于人因与工效学学会年会论文集，第 61 卷。SAGE Publications Sage CA：加利福尼亚州洛杉矶，1370–1374。[17] Spencer C Castro、Helia Hosseinpour、P Samuel Quinan 和 Lace Padilla。2021。利用工作记忆个体差异和 NASA-TLX 考察一维不确定性沟通中的努力程度。（2021）。[18] Spencer C Castro、David L Strayer、Dora Matzke 和 Andrew Heathcote。2019。分散注意条件下认知工作负荷的测量与建模。*Journal of experimental psychology: human perception and performance* 45，6（2019），826。[19] WD Chiles。1982。工作负荷、任务和情境因素作为复杂人类表现的调节因素。人类表现与生产力：压力与表现效能（1982），11–56。[20] W Cole 和 J Davidson。1989。图形表征能够带来快速且准确的贝叶斯推理。载于论文集：计算机应用研讨会

载于《医疗护理》。227–231。[21] William G Cole。1989。通过图形展示理解贝叶斯推理。ACM SIGCHI Bulletin 20, SI（1989），381–386。[22] Andrew RA Conway、Michael J Kane、Michael F Bunting、D Zach Hambrick、Oliver Wilhelm 和 Randall W Engle。2005。工作记忆广度任务：方法学综述与使用指南。Psychonomic Bulletin & Review 12, 5（2005），769–786。[23] Leda Cosmides 和 John Tooby。1996。人类终究是优秀的直觉统计学家吗？重新审视不确定性下判断相关文献中的一些结论。认知 58, 1（1996），1–73。[24] Nelson Cowan、J Scott Saults 和 Lara D Nugent。1997。绝对与相对时间长度在即时记忆遗忘中的作用：以音调高低比较为例。Psychonomic Bulletin & Review 4, 3（1997），393–397。[25] Dick De Waard 和 KA Brookhuis。1996。驾驶员心理负荷的测量。（1996）。[26] David M Eddy。1982。临床医学中的概率推理：问题与机遇。（1982）。[27] Ruth B Ekstrom 和 Harry Horace Harman。1976。《1976年因素参照认知测验套件手册》。教育考试服务中心。[28] Randall W Engle、Michael J Kane、Stephen W Tuholski 等。1999。工作记忆容量的个体差异及其揭示的受控注意、一般流体智力和前额叶皮层功能。载于《工作记忆模型：主动维持与执行控制机制》4（1999），102–134。[29] Jonathan R.G. Etnel、Jasmin M. de Groot、Moad El Jabri、Anouk Mesch、Nathalie A. Nobel、Ad J.J.C. Bogers 和 Johanna J.M. Takkenberg。2020。风险可视化能否提升对数值风险的理解？一项针对总体的随机、研究者设盲问卷调查。《国际医学信息学杂志》

135 2020 104005 10 1016 2019 104005（），。[30] 2013 https://doi.org/ 。/j.ijmedinf. 。Rocio Garcia-Retamero 和 Ulrich Hoffrage。。统计信息的视觉呈现可提升医生及其患者的诊断推断。《社会科学与医学》83（2013），27–33。[31] Gerd Gigerenzer。1994。为何区分单事件概率与频率对心理学（反之亦然）至关重要。载于《主观概率》。Wiley，129–161。[32] Gerd Gigerenzer 和 Ulrich Hoffrage。1995。如何在不进行教学的情况下改进贝叶斯推理：频数格式。《心理学评论》102, 4（1995），684。[33] Amanda L Gilchrist、Nelson Cowan 和 Moshe Naveh-Benjamin。2008。口语句子的工作记忆容量随成年期衰老而下降：老年人回忆的组块数量更少，但组块大小并未变小。《记忆》16, 7（2008），773–787。[34] Sandra G Hart。2006。NASA任务负荷指数（NASA-TLX）：20年后。载于《人因与工效学学会年会论文集》，第50卷。SAGE Publications Sage CA：Los Angeles, CA，904–908。[35] Sandra G Hart 和 Lowell E Staveland。1988。NASA-TLX（任务负荷指数）的制定：实证与理论研究结果。载于《心理学进展》第52卷。爱思唯尔，139–183。[36] Weidong Huang、Peter Eades 和 Seok-Hee Hong。2009。图可视化有效性的测量：认知负荷视角。《信息可视化》8, 3（2009），139–152。[37] Nabil Issa、Mary Schuller、Susan Santacaterina、Michael Shapiro、Edward Wang、Richard E Mayer 和 Debra A DaRosa。2011。应用多媒体设计原则可提升医学教育中的学习效果。《医学教育》45, 8（2011），818–826。[38] Daniel Kahneman。2011。《思考，快与慢》。麦克米伦。[39] Vince Kellen、Susy Chan 和 Xiaowen Fang。2007。借助视觉材料促进条件概率问题的解决。载于《国际人机交互会议》。施普林格，63–71。[40] David O Kennedy 和 Andrew B Scholey。2000。葡萄糖摄入、心率与认知表现：心理努力程度增加的影响。《Psychopharmacology》149, 1（2000），63–71。[41] Azam Khan、Simon Breslav、Michael Glueck 和 Kasper Hornbæk。2015。可视化在乳房X线摄影问题中的益处。《国际人机交互研究杂志》83（2015），94–113。[42] Azam Khan、Simon Breslav 和 Kasper Hornbæk。2018。贝叶斯推断中的交互式教学。《人机交互》33, 3（2018），207–233。[43] M Asif Khawaja、Fang Chen 和 Nadine Marcus。2014。利用语言特征测量认知负荷：对可用性评估和自适应交互设计的启示。《国际人机交互杂志》30, 5（2014），343–368。[44] Robert Kraut、Judith Olson、Mahzarin Banaji、Amy Bruckman、Jeffrey Cohen 和 Mick Couper。2004。在线心理学研究：科学事务委员会互联网研究行为咨询组报告。《美国心理学家》59, 2（2004），105。[45] Elise Lesage、Gorka Navarrete 和 Wim De Neys。2013。进化模块与贝叶斯促进：一般认知资源的作用。《思维与推理》19, 1（2013），27–53。[46] Ottmar V Lipp 和 David L Neumann。2004。持续性操作任务中的注意瞬脱反射调节具有模态特异性。《心理生理学》41, 3

（2004），417–425。[47] Zhengliang Liu、R Jordan Crouser 和 Alvitta Ottley。2020。可视化中个体差异的综述。载于《Computer Graphics Forum》第39卷。Wiley Online Library，693–712。[48] Gerald Matthews、Lauren E Reinerman-Jones、Daniel J Barber 和 Julian Abich IV。2015。心理负荷的心理测量学：多种测量方法虽灵敏，但结果各异。《Human Factors》57, 1（2015），125–143。[49] Luana Micallef、Pierre Dragicevic 和 Jean-Daniel Fekete。2012。通过众包评估可视化对贝叶斯推理的影响。《IEEE可视化与计算机图形学汇刊》18, 12（2012），2536–2545。[50] Tim P Moran。2016。焦虑与工作记忆容量：一项荟萃分析与叙述性综述。《Psychological Bulletin》142, 8（2016），831。[51] Ab Mosca、Alvitta Ottley 和 Remco Chang。2021。交互能否提升借助可视化进行贝叶斯推理的效果？载于《2021年计算机系统中的人为因素CHI会议论文集》。1–14。[52] Lambertus JM Mulder。1992。应用环境中心率与呼吸的测量和分析方法。《Biological Psychology》34, 2-3（1992），205–236。[53] Frederick L Oswald、Samuel T McAbee、Thomas S Redick 和 David Z Hambrick。2015。领域通用工作记忆容量简短测量工具的开发。《Behavior Research Methods》47, 4（2015），1343–1355。[54] Alvitta Ottley、Aleksandra Kaszowska、R. Jordan Crouser 和 Evan M. Peck。2019。文本与可视化结合的奇特案例。载于《EuroVis 2019 - Short Papers》，Jimmy Johansson、Filip Sadlo 和 G. Elisabeta Marai（编）。欧洲计算机图形学协会。https://doi.org/ 10 2312 20191181。/evs。[55] Alvitta Ottley、Blossom Metevier、PK Han 和 Remco Chang。2012。向非专业人士直观传达贝叶斯统计。载于《技术报告》。Citeseer。[56] A. Ottley、E. M. Peck、L. T. Harrison、D. Afergan、C. Ziemkiewicz 和 R. Taylor、H. A.... & Chang。2015。改进贝叶斯推理：措辞、可视化和空间能力的影响。《IEEE可视化与计算机图形学汇刊》22, 1（2015），529–538。[57] Fred Paas、Juhani E Tuovinen、Huib Tabbers 和 Pascal WM Van Gerven。2003。作为推进认知负荷理论手段的认知负荷测量。《教育心理学家》38, 1（2003），63–71。[58] Lace MK Padilla、Spencer C Castro、P Samuel Quinan、Ian T Ruginski 和 Sarah H Creem-Regehr。2019。迈向工作记忆的客观评估：一项使用瞳孔测量和双任务范式的案例研究。《IEEE可视化与计算机图形学汇刊》26, 1（2019），332–342。[59] L. M. Padilla、S. H. Creem-Regehr、M. Hegarty 和 J. K. Stefanucci。2018。借助可视化进行决策：跨学科的认知框架。《Cognitive Research: Principles and Implications》3, 1（2018），1–25。https://doi.org/10.1186/s41235-018-0120-9[60] Harold Pashler。2016。注意。Psychology Press。[61] Evan M M Peck、Beste F Yuksel、Alvitta Ottley、Robert JK Jacob 和 Remco Chang。2013。利用功能性近红外光谱脑感测评估信息可视化界面。载于《计算机系统中的人为因素SIGCHI会议论文集》。473–482。[62] Ulf-Dietrich Reips。2000。网络实验方法：优点、缺点与解决方案。（2000），89–117。[63] Valerie F Reyna。2004。人们如何做出涉及风险的决策：双过程方法。《Current Directions in Psychological Science》13, 2（2004），60–66。[64] Valerie F Reyna。2008。医学决策与健康理论：模糊痕迹理论。《Medical Decision Making》28, 6（2008），850–865。[65] Maria Riveiro、Tove Helldin、Göran Falkman 和 Mikael Lebram。2014。不确定性可视化对目标识别情境中决策的影响。《Computers & Graphics》41（2014），84–98。[66] MS Sanders。1987。McCormick—工程与设计中的人为因素。[67] Valerio Santangelo 和 Emiliano Macaluso。2013。工作记忆对分散注意的贡献。《Human Brain Mapping》34, 1（2013），158–175。[68] Peter Sedlmeier 和 Gerd Gigerenzer。2001。不到两小时教会贝叶斯推理。《Journal of Experimental Psychology: General》130, 3（2001），380。[69] David Spiegelhalter、Mike Pearson 和 Ian Short。2011。对未来的不确定性进行可视化。Science 333, 6048（2011），1393–1400。[70] Nava Tintarev 和 Judith Masthoff。2016。工作记忆个体差异对计划呈现选择的影响。《Frontiers in Psychology》7（2016），1793。[71] Jennifer Tsai、Sarah Miller 和 Alex Kirlik。2011。用于改进贝叶斯推理的交互式可视化。载于《人因与工效学学会年会论文集》第55卷。SAGE Publications Sage CA：Los Angeles, CA，385–389。[72] Amos Tversky 和 Daniel Kahneman。1981。基率的证据影响。技术报告。斯坦福大学心理学系。[73] Amos Tversky 和 Daniel Kahneman。2013。不确定性下的判断：启发式与偏差。载于《金融决策基础手册：第一部分》。World Scientific，261–268。[74] Nash Unsworth、Richard P Heitz、Josef C Schrock 和 Randall W Engle。2005。自动化操作广度任务。《Behavior Research Methods》37, 3（2005），498–505。[75] Patrick Weber、Karin Binder 和 Stefan Krauss。2018年。为什么只有24%的人能用自然频数解决贝叶斯推理问题：尽管存在概率盲视，仍有频数恐惧。*Frontiers in Psychology*9（2018），1833。[76] Lin Yin、Zifu Shi、Zixiang Liao、Ting Tang、Yuntian Xie 和 Shun Peng。2020年。工作记忆和概率格式对贝叶斯推理的影响。*Frontiersin Psychology*

11 2020 863 10 3389 2020 00863 （），。[77] 2016 https://doi.org/。/fpsyg.。Haotian Zhou 和 Ayelet Fishbach。。在网络上进行实验的陷阱：未经关注的选择性流失如何导致令人意外（却错误）的研究结论。人格与社会心理学杂志111，4（2016），493。[78] Bin Zhu 和 Stephanie A Watts。2010年。网络概念的可视化：工作记忆容量差异的影响。信息系统研究21，2（2010），327–344。