基于测验的知识追踪

Shuanghong Shen, Enhon g Chen,高级会员，*IEEE*，Bihan Xu，Qi Liu，Zhenya Huang，Linbo Zhu，Yu S u

arXiv:2304.02413v2 [cs.AI] 6 Apr 2023

摘要——**知识追踪（KT）旨在根据个体在在线学习系统（OIS）中与不同习题的学习交互，评估其不断变化的知识状态。这对于支持后续智能服务（如个性化学习资源推荐）的决策至关重要。现有研究已广泛探讨知识追踪，并提出了许多有效方法。然而，大多数方法都假设学生的历史交互均匀分布在一个连续序列中，忽略了实际交互序列是由一系列边界明确的测验组织而成这一事实：同一测验内的交互连续完成，而不同测验之间的交互则彼此离散，间隔可能长达数天。本文提出基于测验的知识追踪（QKT）模型，根据学生基于测验的学习交互来追踪其知识状态。具体而言，鉴于学生在同一测验内的交互是连续的，且涉及相同或相近的知识概念，我们设计了相邻门控，并在其后接入全局平均池化层，以捕捉测验内短期知识影响。随后，考虑到不同测验通常侧重于不同的知识概念，我们分别采用门控循环单元衡量测验间的知识替代，并采用带有新型近期性感知注意力机制的自注意力编码器衡量测验间的知识互补。最后，我们整合不同测验间长期的知识替代与互补性，以输出学生不断变化的知识状态。在三个公开真实世界数据集上的大量实验结果表明，与现有方法相比，QKT 达到了最先进的性能。进一步分析证实，QKT 有望助力设计更有效的测验。**

索引词—**数据挖掘、神经网络、在线教学系统、知识追踪、基于测验的建模。**

I. 引言

在NLINE学习系统（OIS）在满足个人对智能教育服务日益增长的需求方面发挥着越来越重要的作用 [1, 2],, 例如个性化学习资源推荐 [3, 4]。知识追踪（KT）旨在根据学生的学习情况，监测其不断变化的知识状态

S. Shen、E. Chen（通讯作者）、B. Xu 和 Z. Huang 任职于中国科学技术大学数据科学学院与计算机科学与技术学院、认知智能全国重点实验室及安徽省大数据分析与应用重点实验室，地址：中国安徽省合肥市，邮编 230026。电子邮箱： {closer、xbh0720}@mail.ustc.edu.cn、 {cheneh、huangzhy}@ustc.edu.cn

Q. Liu 任职于安徽省大数据分析与应用重点实验室、中国科学技术大学数据科学学院与计算机科学与技术学院、认知智能全国重点实验室、合肥综合性国家科学中心人工智能研究院，地址：中国安徽省合肥市，邮编 230026。电子邮箱：qiliuql@ustc.edu.cn

Y. Su 任职于中国安徽省合肥市合肥师范学院计算机科学与技术学院及合肥综合性国家科学中心人工智能研究院，邮编：230601。电子邮箱：yusu@hfnu.edu.cn

L. Zhu 任职于中国安徽合肥 230026 合肥综合性国家科学中心人工智能研究院。电子邮箱：lbzhu@iai.ustc.edu.cn

在线教学系统中的交互是提供这些智能服务指导的基础研究任务之一 [5]。近年来，这一新兴研究领域受到越来越多的关注 [6]。

通常，在线教学系统会分配与不同知识概念相关的习题（例如，分数加减法）供学生作答，以帮助他们掌握所需知识。根据学生的交互活动，即他们在不同习题上的表现，研究人员设计了多种知识追踪方法，用于推断学生的知识状态并预测其未来表现。随后，我们可以根据每位学生的知识状态采取有针对性的教学策略，从而提高教学效率。在现有文献中，大多数方法通过序列建模来衡量学生的知识状态。例如，贝叶斯知识追踪（BKT）将学习过程形式化为马尔可夫过程，并利用隐马尔可夫模型评估动态知识状态 [5]。深度知识追踪（DKT）进一步引入循环神经网络/长短期记忆网络 [7] ，对学生学习交互活动进行序列建模 [8]。许多后续研究从不同方面改进了BKT和DKT，例如考虑学生的个体特征 [9, 10, 11],、利用更多辅助信息 [12, 13], 、纳入知识概念的结构 [14, 15]。此外，一些最新研究提出了新的架构来解决知识追踪问题，例如使用记忆网络存储和更新知识状态 [16], ，以及应用注意力机制捕捉学习交互活动中的知识依赖关系 [17, 18]。

然而，包括上述方法在内的大多数现有知识追踪方法，都假设学生的历史交互在一个连续序列中均匀分布，这与实际情况不符。事实上，在线教学系统（OIS）中的习题是以测验的形式分配给学生的，而非逐道单独分配 [19, 20]。具体而言，测验是针对特定知识的一种非正式测试，由若干个知识概念相同或相近的习题组成 [21]。因此，学生的历史交互仅在同一测验内连续，不同测验之间则有明确的边界，它们之间可能间隔数天。为了更好地说明这一点，我们从测验的角度，在图1中展示了一些学生交互序列的真实案例。这些案例取自 Eedi [22], 的真实学习数据。Eedi 是一个在线教学系统，全球数百万名学生每天都会在其中进行交互。在图1中，我们按天记录了学生的交互，并将交互序列开始的第一天标记为起点。图1下半部分清楚地表明，学生的历史交互以测验为单位，且彼此离散。此外，图1上半部分详细展示了s_1在某个测验中连续发生的交互：整个测验作答过程约持续了11分钟

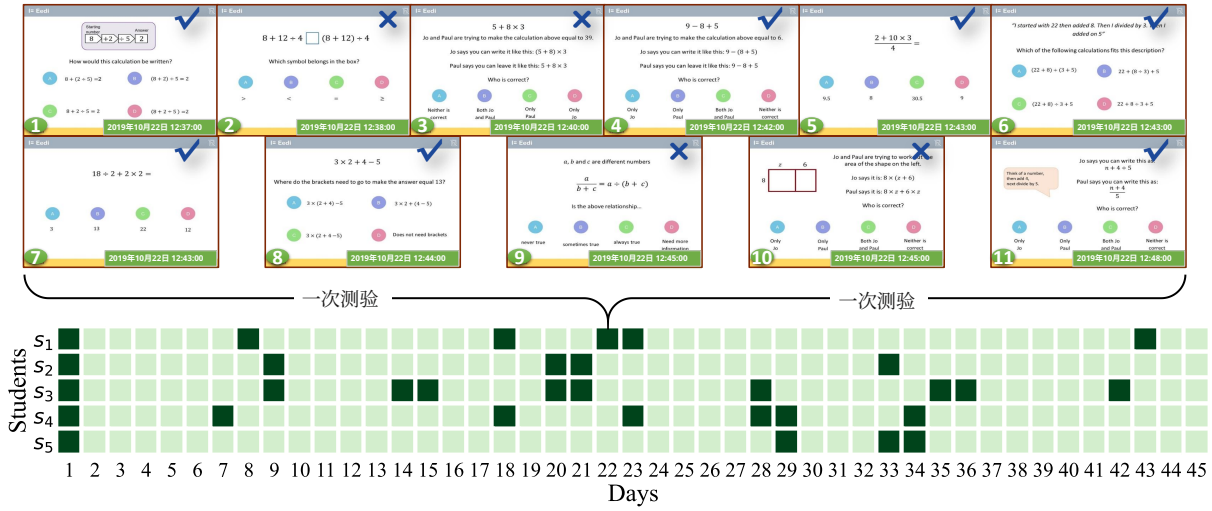


图 1. 从测验的角度展示学生交互序列的真实示例。在下方的框图中，我们按天记录学生的交互，并以第一天作为序列的起点。这里，方框表示一天；深色方框表示学生在相应日期完成了一项或多项测验，浅色方框则表示他们没有作答任何习题。下半部分表明，学生交互序列是按照一系列边界清晰的测验组织的。上半部分展示了 s_1 在某个测验中的连续交互情况：该测验包含 11 道与同一知识概念相关的不同习题：基础算术。 s_1 花了约 11 分钟完成这次测验。

而且，该测验中的所有习题都对应同一个知识概念：基础算术。总而言之，学生在同一测验中的学习交互会在短时间内连续发生，而不同测验之间的交互则彼此离散，间隔一定时间。

本文认为，在知识追踪中考虑学生学习交互活动以测验为基础的组织方式至关重要且大有裨益。遗憾的是，沿着这一方向仍有许多技术和领域方面的挑战需要解决。首先，正如上文所述，同一测验中的习题通常涉及相似的知识概念，学生与之相关的交互会在较短时间内连续发生，因此，如何捕捉测验内短期知识影响并非易事。例如，与一道较难的前序习题相比，作答一道较简单的前序习题，对学生在同一测验中当前习题表现的影响应有所不同[23]。涉及不同测验时，挑战则更多，因为不同测验彼此离散，且通常涉及不同的知识概念。具体而言，如果最近一次测验涉及的知识概念与先前测验相似，那么学生在先前相似测验中的交互可能会变得不可靠，并被近期交互所取代。如何捕捉这种测验间长期知识替代？此外，如果最近一次测验包含先前测验中从未出现过的新知识概念，该如何整合涉及不同知识概念的这些测验，即如何衡量测验间长期知识互补性？

为实现基于测验的知识追踪并应对上述挑战，本文提出基于测验的知识追踪（QKT）模型，通过探索学生基于测验的学习交互来衡量其知识状态。具体而言，我们首先设计了相邻门控，用于控制同一测验中相邻交互之间的知识影响。考虑到学生在一次测验中的平均表现能够反映其在该测验相关知识概念（KC）上的知识状态，我们进一步对每次测验进行全局平均池化。例如，

图1中的学生 s_1 在与同一知识概念（KC）相关的11道习题中答对了7道、答错了4道：基础算术， s_1在基础算术方面的知识状态应约为7/11。随后，我们直接利用门控循环单元（GRU） [24] 评估测验间长期知识替代效应；该单元会记忆学生近期测验中的交互，并遗忘较早的交互。此外，我们提出自注意力编码器，以衡量测验间的知识互补性，从而保留并融合学生在所有历史测验中的交互。为衡量学生在历史测验中知识损失程度的差异，我们在自注意力编码器中设计了一种新颖的近期性感知注意力机制。最后，通过结合不同测验之间的测验间长期知识替代效应与互补性，我们可以获得学生不断演变的知识状态。

我们的主要贡献总结如下：

•我们首先聚焦于在线教学系统（OIS）中面向知识追踪任务的学生学习交互活动的基于测验的组织方式。我们总结了基于测验的交互序列的特点：在单次测验内，交互序列在短时间内连续发生；而在不同测验之间，则以一定间隔离散发生。我们还从测验的角度，对第 VI-A 节中介绍的三个来自不同在线教学系统的真实公开数据集进行了详细分析。

• 我们提出了一种新颖的基于测验的知识追踪模型（QKT），通过探索学生的基于测验的交互序列来评估其动态知识状态。在 QKT 中，我们分别衡量测验内短期知识影响，以及测验间长期知识替代效应和互补性。

• 我们开展了大量实验来验证 QKT 的有效性，结果表明，QKT 的性能优于现有方法。进一步分析表明，QKT 可用于帮助设计更有效的测验。

II. 相关工作

本节将从两个类别介绍现有的相关工作：知识追踪和认知诊断。

A. 知识追踪

随着 OIS 的发展，监测学生知识状态的重要性日益凸显[6]，，Corbett 和 Anderson 首次将其正式定义为知识追踪任务 [5]。他们提出了 BKT 模型，将学习过程视为马尔可夫过程，并利用学生观察到的交互序列推断其潜在知识状态。随后，研究人员从多个方面丰富并发展了 BKT。例如，为每位学生单独设置 BKT 参数 [9, 10]，，考虑 OIS 中的导师干预 [25]，，以及纳入学生的遗忘效应 [26]。近年来，深度学习（DL）的进展推动了基于神经网络的知识追踪模型发展。具体而言，DKT 引入循环神经网络/长短期记忆网络，以序列方式建模学生的知识状态 [8]。随后，DKVMN 使用记忆网络存储和更新学生在特定知识概念上的潜在知识状态 [16]。一些研究人员考虑到知识概念内部的自然结构，并提出使用图神经网络来捕捉不同知识概念之间知识状态的影响整合 [27, 14, 15]。此外，一些研究注意到，学生历史上相关的交互对其未来表现影响更大。因此，他们引入注意力机制来建模学习中的知识依赖关系。例如，Pandey 和 Karypis[17] 应用 Transformer [28] 追踪学生的知识状态；Ghosh 等人 [18] 将自注意力机制与单调性假设结合。Zhang 等人 [29] 应用双重注意力机制，基于多个因素建模学生的学习进程。还有许多研究聚焦于习题表示。例如，从习题文本内容中学习其语义表示 [30, 31]。Liu 等人 [32]则尝试基于习题与知识概念的关系、习题相似度、知识概念相似度以及习题难度，共同学习预训练习题嵌入。Shen 等人 [33] 对习题难度效应进行建模，并设计自适应序列神经网络，使习题难度与知识状态相匹配。近期，研究人员进一步探索了学生的学习过程。Wang 等人 [34]提出使用霍克斯过程，自适应地建模学习中的时间交叉效应。Shen 等人 [35]提出对学生的学习增益和遗忘进行建模，以计算其动态知识状态。Long 等人 [36]估计了学生的个体认知水平以及学习过程中的知识获取。

总之，现有的大多数知识追踪方法遵循序列建模范式。它们假设学生的历史交互序列在连续序列中均匀分布，忽略了学生的交互以测验为基础且具有明确边界这一事实。因此，不同测验之间的交互实际上是离散的。尽管一些研究注意到了交互时间戳的重要性 [37, 35]，，但它们仅将时间信息作为附加特征使用。Ke 等人 [38]进一步根据固定时长将学生的历史交互划分为多个会话，并采用会话感知知识追踪。然而，这种做法同样不符合学生交互以测验为基础的组织方式，并损害了测验内及测验间的知识相关性。

B. 认知诊断

认知诊断（CD）同样关注根据个体的行为评估其知识状态 [39]。与知识追踪（KT）不同，认知诊断通常应用于测试场景，会利用所有历史交互来学习每个学生的静态知识状态。具体而言，项目反应理论（IRT）是经典的认知诊断模型之一 [40]，，它使用逻辑回归模型来估计学生的知识状态：

$$p = c + \frac{1-c}{1+e^{-(\theta-\beta)}}, \tag{1}$$

其中，c 是随机猜测概率，θ 是知识状态，β 是练习难度。然而，IRT 中的 θ 是单一值，无法反映学生在不同知识概念（KCs）上的知识状态。因此，研究者提出了多维项目反应理论（MIRT），使用多维向量表示学生在不同知识概念（KCs）上的知识状态[41，42]。近年来，深度学习已广泛应用于认知诊断 [43, 44, 45]。例如，神经认知诊断（NCD）尝试利用神经网络对学生-练习交互进行建模。NCD 的通用框架可表示为：

$$p = \phi_n(...\phi_1(F^s, F^{kc}, F^{other}, \theta_f)), \tag{2}$$

其中，ϕ 表示用于建模学生与习题交互的神经网络。F^s 表示知识状态，F^{kc} 表示知识成分因子，F^{other} 表示其他因素（如练习难度），θ_f 表示所有可学习参数。随后，研究人员从不同方面扩展了 NCD，例如纳入学生丰富的上下文信息[46]，以及衡量学生、习题和知识概念之间的层级关系 [47]。然而，认知诊断（CD）隐含地假设学生的所有交互对其知识状态都同等重要。在单次测试/测验中，这一假设是合理的，但对于现实中学生基于测验的交互序列而言，未必可靠。

III. 问题陈述

OIS包含多个基本元素，包括学生、习题、知识概念以及学生-练习交互。假设数据集中所有学生共同构成学生集合，即 $\mathbb{S} = \{s_1, s_2, ..., s_S\}$，所有习题共同构成练习集合，即 $\mathbb{E} = \{e_1, e_2, ..., e_E\}$，所有知识概念共同构成知识成分集合，即 $\mathbb{K} = \{k_1, k_2, ..., k_K\}$，其中$s$、$e$、$k$分别表示一名学生、一道习题和一个知识概念，$S$、$E$、$K$分别表示学生、习题和知识概念的数量。一般来说，每道习题都与特定的知识概念相关，我们使用由教育专家给出的Q矩阵来表示习题与知识概念的关系。Q矩阵由1和0组成，其中1表示对应的习题与知识概念相关，否则为0。对于特定学生

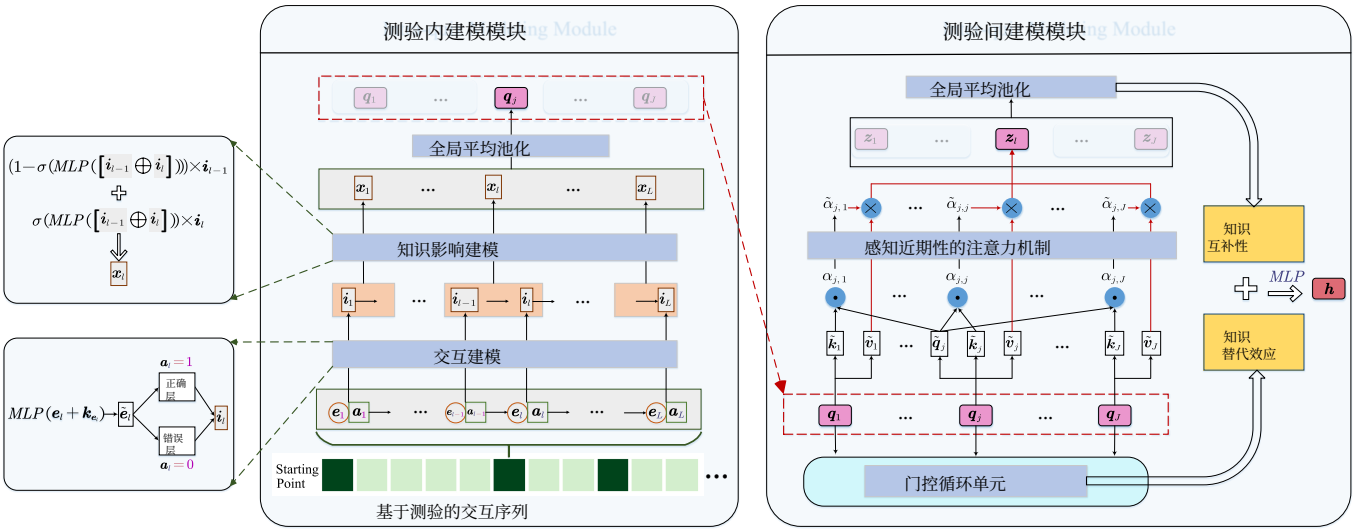


图 2. QKT 模型的架构。

s，学生-练习交互是最基本的单元。我们有 i 表示 s的一次交互，其中包括一道习题e 以及 s 在 e上给出的答案 a，即 $i = (e、\ a|s)$。这里答案 a 是一个二元正确性标签（1表示正确，0表示错误）。测验是OIS中习题的基本组织形式。一般来说，测验表现为对特定知识的非正式测试，由具有相同或相似知识概念的多道习题组成。例如，图1中给出的测验包含11道具有相同知识概念的习题：基础算术。需要注意的是，测验在不同系统中可能有不同的名称，例如在ASSISTments [48] 和CodeWorkout [49]中被称为作业。为方便起见，我们在本文中统一使用测验这一名称。我们有 q 表示学生在一次测验上的交互，即 $q = \{i_1, i_2, ..., i_L\}$，$L$是测验 q的长度（即 q中交互的长度），下标从1到 L 表示每次交互的顺序。令 U 表示学生的整个交互序列，我们可以将 U表示为测验集合，即 $U = \{q_1, q_2, ..., q_J\}$，$J$是测验的数量，下标从1到 J 表示每次测验的顺序。然后，我们可以将基于测验的知识追踪任务正式表述如下：

a) 问题形式化。：给定一名学生在多个测验中的序列交互记录 $U = \{q_1, q_2, ..., q_J\}$，基于测验的知识追踪任务旨在评估该学生在不同测验中的动态知识状态，并预测其在未来测验中新习题上的表现。

IV. 基于测验的知识追踪

本节将详细介绍所提出的 QKT 模型，并说明如何衡量测验内短期知识影响，以及测验间长期知识替代和互补性。QKT 的架构如图 2 和算法 1 所示，主要由测验内建模模块和测验间建模模块组成。具体而言，在测验内建模模块中，我们主要关注探索学生在一次测验中相邻交互之间的知识影响，并捕捉他们每次测验的整体知识状态。随后，在测验间建模模块中，我们转而衡量不同测验之间的知识替代和知识互补，并最终将二者整合，以输出学生在不同测验中的知识状态变化。

A. 测验内建模模块

在此模块中，给定学生在测验$q = \{i_1, i_2, ..., i_L\}$中的交互记录，我们旨在衡量测验内短期知识影响，并输出测验向量 $\boldsymbol{q}$，以表示该学生在测验 q中的知识状态。

1) 交互建模：一个测验通常包含多个交互，每个交互由习题-作答 对组成。因此，我们首先在此部分进行交互建模。

具体而言，我们使用嵌入矩阵 $\boldsymbol{E} \in \mathbb{R}^{E \times d_e}$ 表示数据集中的所有习题。因此，对于测验 q中的特定习题 e_l，我们可以从$\boldsymbol{E}$中获取其嵌入 $\boldsymbol{e}_l$。此外，考虑到习题涉及不同的知识概念，我们还使用嵌入矩阵 $\boldsymbol{K} \in \mathbb{R}^{K \times d_k}$（此处设定 d_e 等于 d_k）来表示所有知识概念，因此也可以从 $\boldsymbol{K}$中获取 e_l 的知识概念嵌入 $\boldsymbol{k}_{e_l}$。随后，我们通过多层感知机（MLP）将习题嵌入 $\boldsymbol{e}_l$ 与其知识概念嵌入 $\boldsymbol{k}_{e_l}$ 结合起来，以表示完整习题：

$$\tilde{\boldsymbol{e}}_l = \text{Relu}(\boldsymbol{W}_1^T(\boldsymbol{e}_l + \boldsymbol{k}_{e_l}) + \boldsymbol{b}_1), \tag{3}$$

其中，$\tilde{\boldsymbol{e}}_l \in \mathbb{R}^{d_e}$ 表示完整习题向量，$\boldsymbol{W}_1 \in \mathbb{R}^{d_e \times d_e}$ 和 $\boldsymbol{b}_1 \in \mathbb{R}^{d_e}$ 是可训练参数。

随后，考虑到二元答案值，我们分别提出正确层和错误层，以区分两种二元答案的不同影响。然后，习题-答案 对可表示如下：

$$\boldsymbol{i}_l = \begin{cases} \boldsymbol{W}_r^T \tilde{\boldsymbol{e}}_l + \boldsymbol{b}_r, & \text{if} \quad a_l = 1, \\ \boldsymbol{W}_w^T \tilde{\boldsymbol{e}}_l + \boldsymbol{b}_w, & \text{if} \quad a_l = 0, \end{cases} \tag{4}$$

其中，$\boldsymbol{i}_l \in \mathbb{R}^{d_i}$ 表示交互向量。$\boldsymbol{W}_r \in \mathbb{R}^{d_e \times d_i}$ 和 $\boldsymbol{b}_r \in \mathbb{R}^{d_i}$ 是正确层的可训练参数，$\boldsymbol{W}_w \in \mathbb{R}^{d_e \times d_i}$ 和 $\boldsymbol{b}_w \in \mathbb{R}^{d_i}$ 是错误层的可训练参数。

2) 知识影响建模：完成交互建模后，我们继续测量测验中各交互之间的知识影响。具体而言，这种知识影响主要存在于多个方面的相邻交互中。例如，完成上一道习题所花费的精力不同，对学生完成当前习题的表现会产生不同影响 [50]。此外，先前的难题比简单题带来更多负面影响；也就是说，先简单后较难的习题会产生学习效应，而按由难到易的顺序完成习题则会产生疲劳效应 [51, 23]。

在我们提出的 QKT 模型中，我们以统一的方式衡量上述相邻的多方面知识影响。具体而言，我们根据两个相邻交互之间的影响设计相邻门控，用于控制分别从前一交互和当前交互中提取多少信息。计算公式如下：

$$\begin{aligned} \boldsymbol{\Gamma}_l &= \sigma(\boldsymbol{W}_2^T(\boldsymbol{i}_{l-1} \oplus \boldsymbol{i}_l)) + \boldsymbol{b}_2, \\ \boldsymbol{x}_l &= \boldsymbol{\Gamma}_l \cdot \boldsymbol{i}_l + (1 - \boldsymbol{\Gamma}_l) \cdot \boldsymbol{i}_{l-1}, \end{aligned} \tag{5}$$

其中，$\boldsymbol{\Gamma}_l \in \mathbb{R}^{d_i}$ 表示相邻门控，$\boldsymbol{x}_l \in \mathbb{R}^{d_i}$表示两个相邻交互的组合向量。$\sigma$ 是*sigmoid* 激活函数，$\cdot$ 表示逐元素乘积运算，$\oplus$ 表示向量拼接，$\boldsymbol{W}_2 \in \mathbb{R}^{2d_i \times d_i}$ 和$\boldsymbol{b}_2 \in \mathbb{R}^{d_i}$ 是可训练参数。

3）全局平均池化：在这一部分，我们进一步探讨如何权衡测验中的所有交互，以获得测验向量，用来表示学生在该测验中的整体知识状态 q。由于测验中的交互涉及相同或相似的知识概念，每个交互都应对整体结果有所贡献。因此，我们对 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L\}$ 执行全局平均池化运算，以计算测验向量：

$$\boldsymbol{q} = \sum_{l=1}^{L} (\boldsymbol{x}_l)/L. \tag{6}$$

其中，$\boldsymbol{q} \in \mathbb{R}^{d_q}$ （d_q 等于 d_i）表示测验向量。通过上述建模过程，$\boldsymbol{q}$ 同时包含测验内短期知识影响，以及学生在与测验相关的知识概念上的整体知识状态。

B. 测验间建模模块

在此模块中，我们将关注点从一次测验转向多次测验，并尝试衡量测验间的长期知识整合，即将不同的测验向量整合在一起，以表示学生动态变化的知识状态。具体而言，测验间长期知识整合主要有两种形式：知识替代和知识互补。我们将先分别评估这两种形式，再将它们结合起来。

1) 知识替代：知识替代是指学生在先前测验上的交互会被近期交互所取代，这通常出现在具有相同或相似知识概念的测验中。例如，学生在刚开始学习知识概念维恩图时，可能在相关测验上表现较差。

算法 1QKT 模型。

输入：学生的基于测验的交互序列，$U = \{q_1, q_2, ..., q_J\}$；学生在每个测验中的交互，$q = \{i_1, i_2, ..., i_L\}$；每次交互的习题 e 以及学生的作答 a；指示习题与知识概念关系的Q矩阵。**输出：**学生的知识状态，$\boldsymbol{h} \in \mathbb{R}^{d_h}$。1: 通过公式(3)计算学生交互中所有习题的完整习题向量；2: 通过公式(4)获得学生所有交互的交互向量；3: 通过公式(5)计算相邻门控；4: 通过公式(6)获得学生所有已完成测验的测验向量；5: 通过公式(7)度量测验间知识替代；6: 通过公式(8)、公式(9)计算近期性感知注意力值；7: 通过公式(10)、公式(11)度量测验间知识互补性；8: 通过公式(12)整合学生的知识状态；9: **返回** $\boldsymbol{h}$

然而，经过一段时间的学习后，他在后续测验中能够取得满分，因此，他在最初几次测验中的交互并不可靠，需要根据近期交互进行更新。

为衡量上述知识替代效应，我们直接使用门控循环单元（GRU） [24] 对学生的测验序列 $\{\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_J\}$ 进行建模，如下所示：

$$\begin{aligned} \boldsymbol{\Gamma}_r &= \sigma(\boldsymbol{W}_3^T(\boldsymbol{sub}_{j-1} \oplus \boldsymbol{q}_j) + \boldsymbol{b}_3), \\ \boldsymbol{\Gamma}_u &= \sigma(\boldsymbol{W}_4^T(\boldsymbol{sub}_{j-1} \oplus \boldsymbol{q}_j) + \boldsymbol{b}_4), \\ \tilde{\boldsymbol{sub}}_j &= tanh(\boldsymbol{\Gamma}_r \cdot \boldsymbol{W}_5^T(\boldsymbol{sub}_{j-1} \oplus \boldsymbol{q}_j) + \boldsymbol{b}_5), \\ \boldsymbol{sub}_j &= (1 - \boldsymbol{\Gamma}_u) \cdot \boldsymbol{sub}_{j-1} + \boldsymbol{\Gamma}_u \cdot \tilde{\boldsymbol{sub}}_j, \end{aligned} \tag{7}$$

其中，$\boldsymbol{sub}_{j-1} \in \mathbb{R}^{d_q}$ 表示所有先前测验的汇总，$\boldsymbol{\Gamma}_r$ 是重置门，用于确定如何结合当前测验向量与先前记忆 $\boldsymbol{sub}_{j-1}$；$\boldsymbol{\Gamma}_u$ 是更新门，用于控制保留多少先前记忆。$tanh$ 是激活函数，$\boldsymbol{W}_3 \in \mathbb{R}^{2d_q \times d_q}$、$\boldsymbol{W}_4 \in \mathbb{R}^{2d_q \times d_q}$、$\boldsymbol{W}_5 \in \mathbb{R}^{2d_q \times d_q}$、$\boldsymbol{b}_3 \in \mathbb{R}^{d_q}$、$\boldsymbol{b}_4 \in \mathbb{R}^{d_q}$、$\boldsymbol{b}_5 \in \mathbb{R}^{d_q}$ 是可训练参数。按顺序处理所有测验后，我们可以得到向量 $\boldsymbol{sub}_J$，该向量捕捉了测验间长期知识替代：它对近期测验中的更多学生交互进行记忆，并更多地遗忘远期测验。

2）知识互补性：与知识替代不同，知识互补性意味着，学生在先前测验中的交互会与其在后续测验中的交互并行发挥作用。具体而言，在学生完成的所有测验中，我们针对知识概念重叠的测验衡量了知识替代效应。然而，还有许多测验关注不同的知识概念，学生在这些测验中的交互应相互补充，从而更全面地检索知识状态。

为衡量上述知识互补性，我们采用自注意力编码器来捕捉不同测验之间的依赖关系。具体而言，对于每个测验向量 $\boldsymbol{q}_j$，我们首先利用三个嵌入层，分别将 $\boldsymbol{q}_j$ 映射为查询向量 $\tilde{\boldsymbol{q}}_j \in \mathbb{R}^{d_q \times 1}$、键向量 $\tilde{\boldsymbol{k}}_j \in \mathbb{R}^{d_q \times 1}$ 和值向量 $\tilde{\boldsymbol{v}}_j \in \mathbb{R}^{d_q \times 1}$。然后，计算 $\alpha_{jj'}$ 与 $\boldsymbol{q}_j$ 和 $\boldsymbol{q}_{j'}$ 之间的点积注意力值：

$$\alpha_{jj'} = \text{softmax}(\tilde{\boldsymbol{q}}_j \cdot \tilde{\boldsymbol{k}}_{j'}) = \frac{\exp(\tilde{\boldsymbol{q}}_j \cdot \tilde{\boldsymbol{k}}_{j'})}{\sum_{j'=1}^{J} \exp(\tilde{\boldsymbol{q}}_j \cdot \tilde{\boldsymbol{k}}_{j'})}. \quad (8)$$

然而，$\alpha_{jj'}$ 仅将 $\boldsymbol{q}_j$ 和 $\boldsymbol{q}_{j'}$之间的相似度作为其注意力权重，却没有考虑特定测验在整个测验序列中的顺序。实际上，学习具有时间性，近期事件对学生的影响更大，这也称为近因效应 [52]。因此，我们认为有必要赋予近期测验更高的权重，并进一步提出近期性感知注意力机制，在注意力值 $\alpha_{jj'}$ 中加入两个近期性感知项，如下所示：

$$\begin{aligned}
\beta^1_{j'} &= \text{softmax}(\gamma j') = \frac{\exp(\gamma j')}{\sum_{j'=1}^{J} \exp(\gamma j')}, \\
\beta^2_{j'} &= \text{softmax}(\gamma(J - j')) = \frac{\exp(\gamma(J - j'))}{\sum_{j'=1}^{J} \exp(\gamma(J - j'))}, \\
\tilde{\alpha}_{jj'} &= \alpha_{jj'} + \beta^1_{j'} - \beta^2_{j'},
\end{aligned} \quad (9)$$

其中，γ 是一个常数参数，用于缩放近期性感知项 $\beta^1_{j'}$ 和 $\beta^2_{j'}$，使其与点积注意力值 $\alpha_{jj'}$相匹配。需要注意的是，通过加上 $\beta^1_{j'}$ 并减去 $\beta^2_{j'}$，我们只是将部分点积注意力值从学生测验序列中较早的测验转移到较晚的测验，从而实现近期测验更重要这一假设。点积注意力的总值保持不变，即$\sum_{j'=1}^{J} \tilde{\alpha}_{jj'} = \sum_{j'=1}^{J} \alpha_{jj'} = 1$。

随后，将 $\tilde{\alpha}_{jj'}$ 与 $\tilde{\boldsymbol{v}}_j$ 相乘，得到由各值加权求和构成的输出：

$$\boldsymbol{z}_j = \sum_{j'=1}^{J} \tilde{\alpha}_{jj'} \tilde{\boldsymbol{v}}_{j'}, \quad (10)$$

其中，$\boldsymbol{z}_j \in \mathbb{R}^{d_q \times 1}$ 表示 q_j的测验向量，其中包含与其他测验的知识互补性。最后，与第 IV-A3 节中测验内交互的整合方式类似，我们对 $\{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_J\}$ 执行全局平均池化：

$$\boldsymbol{com}_J = \sum_{j=1}^{J} (\boldsymbol{z}_j)/J. \quad (11)$$

然后，我们可以得到向量 $\boldsymbol{com}_J$，它捕捉了测验间长期知识互补性。

3) 整合知识状态：在计算出分别用于建模测验间长期知识替代和互补性的 $\boldsymbol{sub}_J$和 $\boldsymbol{com}_J$ 后，我们需要进一步将二者结合起来，以输出学生在不同测验中的知识状态。在 QKT 中，我们采用一种简单而有效的加法方式来整合它们：

$$\boldsymbol{h} = \boldsymbol{W}_6^T(\boldsymbol{sub}_J + \boldsymbol{com}_J) + \boldsymbol{b}_6, \quad (12)$$

其中，$\boldsymbol{h} \in \mathbb{R}^{d_h}$ 表示知识状态向量，$\boldsymbol{W}_6 \in \mathbb{R}^{d_q \times d_h}$ 和 $\boldsymbol{b}_6 \in \mathbb{R}^{d_h}$ 是可训练参数。

V. 模型学习

为了训练 QKT 中的所有嵌入、权重矩阵和偏置项，我们首先使用 $\boldsymbol{h}$ 预测学生在未来测验中对习题 e_n 的作答，然后选择预测答案 y_n 与学生实际答案 a_n 之间的交叉熵对数损失作为目标函数，如下所示：

$$\begin{aligned}
y_n &= \sigma(\boldsymbol{W}_8^T(\boldsymbol{W}_7^T(\tilde{\boldsymbol{e}}_n \cdot \boldsymbol{h} \oplus \tilde{\boldsymbol{e}}_n \oplus \boldsymbol{h}) + \boldsymbol{b}_7) + \boldsymbol{b}_8), \\
\mathbb{L}(\theta) &= -\sum (a_n \log y_n + (1 - a_n) \log(1 - y_n)) + \lambda_\theta ||\theta||^2,
\end{aligned} \quad (13)$$

其中，我们同时使用乘法（即 $\tilde{\boldsymbol{e}}_n \cdot \boldsymbol{h}$）和拼接（即 $\tilde{\boldsymbol{e}}_n \oplus \boldsymbol{h}$）来整合预测向量；$\boldsymbol{W}_7 \in \mathbb{R}^{d_h \times d_h}$、$\boldsymbol{W}_8 \in \mathbb{R}^{d_h \times 1}$、$\boldsymbol{b}_7 \in \mathbb{R}^{d_h}$、$\boldsymbol{b}_8 \in \mathbb{R}^1$ 是可训练参数。θ 表示 QKT 中的所有可训练参数，λ_θ 是正则化项 c。目标函数将使用 Adam优化器 [53]在小批量上进行最小化。

VI. 实验

在本节中，我们首先介绍实验中使用的公开真实世界数据集。然后，我们开展实验以评估 QKT 的有效性，旨在回答以下研究问题：

- **RQ1:** 我们提出的 QKT 模型在学生表现预测任务上是否优于现有方法?
- **RQ2:** QKT 中的不同组件分别如何影响其性能?
- **RQ3:** 每位学生测验数量的变化对 QKT 有何影响?
- **RQ4:** 每个测验交互长度的变化对 QKT 有何影响?

A. 数据集

我们的实验使用三个公开真实世界数据集进行评估：(1) Assist2012[1]；(2) Eedi2020[2]；(3) CSEDM[3]。所有数据集的统计数据见表I。图3还展示了各数据集中测验长度 L 和测验数量J 的分布，这些分布在不同数据集之间各不相同。所有数据集的详细说明如下：

•**Assist2012** 来自 ASSISTments 数学辅导系统 [48] 中2012-2013学年八年级学生的数据。ASSISTments 中知识概念相似的习题被组织成作业 （类似于测验），学生需要完成不同的作业 来掌握相关知识。在我们的实验中，我们过滤掉了缺少习题相关知识概念的交互数据。

[1] https://sites.google.com/site/assistmentsdata/datasets/2012-13-school-data-with-affect [2] https://eedi.com/projects/neurips-education-challenge [3] https://sites.google.com/ncsu.edu/csedm-dc-2021/home

表I 所有数据集的统计数据。

统计数据	数据集		
	Assist2012	Eedi2020	CSEDM
学生人数	29,018	4,918	840
习题数量	53,091	948	50
知识概念数量	265	53	50
交互次数	2,711,813	1,382,727	38,531
每名学生的平均交互次数	93.45	281.16	45.87
每名学生的平均测验次数	19.02	20.54	4.73
每个测验的平均交互次数	4.93	13.74	9.70

•**Eedi2020** 发表于 NeurIPS 2020 教育挑战赛，包含学生对 Eedi 数学问题的作答数据。Eedi 是一个在线教学系统，从 2018 学年到 2020 年，每天都有全球数百万学生与之交互。Eedi 中的习题被组织成不同的测验。我们使用了该挑战赛中任务 3 和 4 的数据。该数据集包含层级化的知识概念，我们仅使用每道习题对应的叶节点知识概念。

• **CSEDM** 发表于第二届计算机科学教育数据挖掘挑战赛。该数据集收集自美国一所公立大学 2019 年春季和秋季学期的 CS1 课程，包含学生针对 5 个不同作业中的 50 道编程题提交的代码。

B. 实验设置

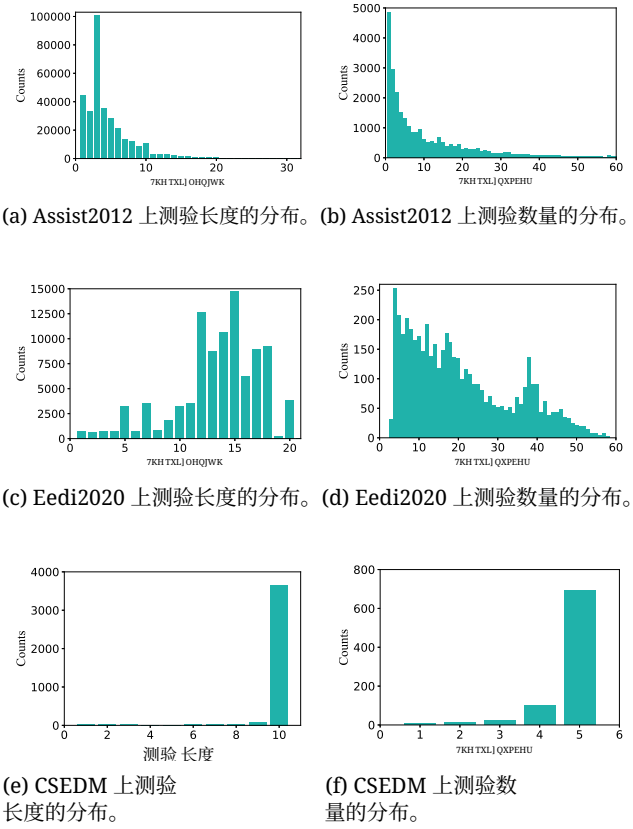
对于所有学生的作答记录，我们首先按作答时间戳对其排序。然后，我们利用学生在之前所有 $J-1$ 测验中的交互来训练模型，并预测他们在最后一次测验（即第 J 次测验）中习题的表现。为确保实验结果的可靠性，我们筛除了作答测验次数少于 2 次的学生。为设置训练过程，我们从均匀分布 [54]中随机初始化所有参数和嵌入。所有数据集上的实验均按学生进行 5 折交叉验证[4]。初始学习率为 1e-3，并设置每三个周期将学习率衰减 50%，以达到最优点。小批量大小为 32。维度d_e、 d_i、 d_q和 d_h 均统一设为 128。用于缩放公式（9）中近期性相关项取值的常数参数 γ 为 1e-5。

根据如图3所示所有数据集上测验长度和测验数量的分布，对于Assist2012，在我们的实验中测验长度和测验数量均为30。对于Eedi2020，我们分别将测验长度和测验数量设置为20和50。对于CSEDM，测验长度为10，测验数量为5。

C. 对比基线方法

为了验证QKT的有效性，我们将其与现有的知识追踪方法进行了比较。所有对比方法都经过调优至
以确保比较公平，取得最佳性能。所有模型均使用 TensorFlow 实现，并在由多台配备 NVIDIA Tesla V100 GPU 的

[4]我们将在录用后公开代码



(a) Assist2012 上测验长度的分布。(b) Assist2012 上测验数量的分布。

(c) Eedi2020 上测验长度的分布。(d) Eedi2020 上测验数量的分布。

(e) CSEDM 上测验长度的分布。

(f) CSEDM 上测验数量的分布。

图 3. 所有数据集上测验长度和测验数量的分布。

Linux 服务器组成的集群上进行训练。知识追踪基线方法的简要介绍如下：

• **深度知识追踪** 引入循环神经网络/长短期记忆网络，以顺序方式对学生的知识状态进行建模 [8]。
• **DKVMN** 使用记忆网络来存储和更新知识状态 [16]。

• **SAKT** 利用自注意力机制捕捉学生与习题交互之间的知识依赖关系 [17]。

• **AKT** 通过自注意力编码器学习上下文感知的交互表示，并衡量学生过去获得的、与当前习题相关的知识 [18]。

• **LPKT** 对学习过程进行建模，并计算学生的学习增益和遗忘情况，以评估其知识状态[35]。

• **DIMKT** 衡量练习难度对知识状态和学习过程的影响 [33]。

此外，我们还将 QKT 与认知诊断领域的两种基线方法进行比较：
• **项目反应理论** 使用逻辑函数将学生的知识状态建模为连续变量 [40]。
• **NCD** 采用神经网络来学习复杂的学生-习题交互 [45]。

D. 评估指标

为评估 QKT 和所有基线方法的性能，我们从回归和分类两个角度采用多种指标

表II 学生表现预测中所有对比方法的结果（$\times 100$）。我们用下划线标出各列中所有基线方法的最佳结果，粗体数字则表示整列中的最佳结果。 ± 之后的误差线表示各方法 5 次评估运行的标准差。增益表示我们的模型相较于所有基线方法中最佳表现的提升。

	Assist2012			Eedi2020			CSEDM		
方法	AUC	RMSE	r^2	AUC	RMSE	r^2	AUC	RMSE	r^2
IRT	70.91±0.34	44.98±0.17	12.02±0.29	68.67±0.67	47.22±0.20	10.78±0.77	72.54±1.67	24.40±0.93	2.32±1.71
NCD	70.66±0.27	45.25±0.30	10.99±0.75	69.91±0.47	47.18±0.23	10.94±0.88	74.36±1.38	24.13±1.05	4.16±0.41
DKT	70.28±0.57	45.08±0.28	11.63±0.78	73.31±0.65	45.77±0.26	16.20±0.94	83.47±2.58	21.76±1.70	12.44±6.49
DKVMN	64.36±0.66	47.61±0.67	1.45±2.37	73.21±0.50	45.78±0.23	16.14±0.81	83.89±2.56	21.92±1.61	8.82±6.91
SAKT	66.51±0.58	46.57±0.46	5.69±0.46	73.62±0.29	45.52±0.11	17.11±0.38	79.78±4.33	22.31±1.34	8.37±5.35
AKT	70.99±0.62	45.23±0.59	11.03±2.33	74.48±0.38	45.15±0.15	18.43±0.52	84.26±2.21	21.72±1.38	10.40±6.63
LPKT	70.36±0.76	44.99±0.34	11.99±0.94	74.38±0.39	45.33±0.17	17.81±0.60	84.96±2.25	21.49±1.22	13.50±3.43
DIMKT	72.04±0.40	44.53±0.29	13.78±0.79	74.35±0.49	45.35±0.21	17.71±0.74	85.04±1.74	21.50±1.16	13.30±5.22
QKT	**72.71±0.61**	**44.28±0.32**	**14.72±0.83**	**75.23±0.37**	**44.83±0.17**	**19.68±0.59**	**85.40±1.58**	**21.39±1.15**	**14.31±3.63**
Gain	0.93	0.56	6.82	1.00	0.71	6.78	0.42	0.47	6.00

数据集 Assist2012 上消融实验的表III结果（$\times 100$）。

方法	测验内影响	测验间替代效应	测验间互补性	近期性感知注意力	AUC	RMSE	r^2
QKT 不含知识注入	✘	✔	✔	✔	72.50	44.39	14.33
QKT 不含 SUB	✔	✘	✔	✔	71.90	44.56	13.66
QKT 不含 COM	✔	✔	✘	✔	72.49	44.37	14.40
QKT 不含近期性感知注意力	✔	✔	✔	✘	72.62	44.33	14.55
QKT	✔	✔	✔	✔	**72.71**	**44.28**	**14.72**

进行评估。具体而言，从分类任务的角度，我们使用ROC曲线下面积（曲线下面积）衡量效果，数值越大，结果越好。对于回归任务，我们使用均方根误差（均方根误差）和皮尔逊相关系数的平方（r^2）量化预测答案与实际答案之间的差异。均方根误差越小，结果越好。r^2 则相反，数值越大，结果越好。

E. 学生表现预测（*RQ1*）

学生表现预测任务对于评估所捕获知识状态的质量至关重要，即预测越准确，说明对知识状态的估计越准确。为评估 QKT 的有效性，我们将其在此任务上的表现与所有基线方法进行比较。表II展示了我们模型和所有基线方法的实验结果，从中可以得出几项重要发现。

首先，QKT在所有数据集和所有指标上的表现均优于现有方法，这表明我们提出的QKT能够有效捕捉学生学习交互活动的基于测验的组织方式。其次，与一些将作答时间和间隔时间作为额外特征来建模整个学习过程的基线方法（如LPKT）相比，QKT无需使用任何时间信息便取得了更好的表现，这进一步证明了从测验视角探索学生交互序列的重要性和价值。需要指出的是，时间信息对QKT同样重要；由于本文主要关注发现并评估开展基于测验的知识追踪的必要性，我们将QKT扩展以纳入时间信息留待未来工作

。第三，QKT相较于最佳基线方法所取得的性能提升与学生的平均测验数量呈正相关，这符合我们的直觉：测验数量越多，QKT从测验间建模模块中获益越多（我们将在第VI-G节进一步验证这一点）。例如，Assist2012和Eedi2020数据集中学生参加测验的平均数量相近（即Assist2012为19.02，Eedi2020为20.54，如表I所示），QKT相较于最佳基线方法在这两个数据集上取得的性能提升也相近。然而，CSEDM中的学生完成的测验较少，因此QKT的性能提升相对较小。

F. 消融研究（*RQ2*）

本节通过消融实验，展示 QKT 的不同组件如何影响其性能。表III展示了模型在 Assist2012 上的实验结果。我们设计了 QKT 的四种变体，每种变体保留 QKT 中的一个组件，并保留其他组件。具体如下：

- **不含 KI 的 QKT** 指不包含知识影响建模的 QKT，即在不含 KI 的 QKT 中移除了相邻门控及其对应操作。

- **不含 SUB 的 QKT** 指不包含知识替代建模的 QKT，其中移除了门控循环单元模块，并假设测验间仅存在长期知识互补；也就是说，在不含 SUB 的 QKT 中，学生的知识状态完全由 $\boldsymbol{com}_j$ 表示。

- **不含 COM 的 QKT** 指不进行知识互补建模的 QKT，其中我们移除了自

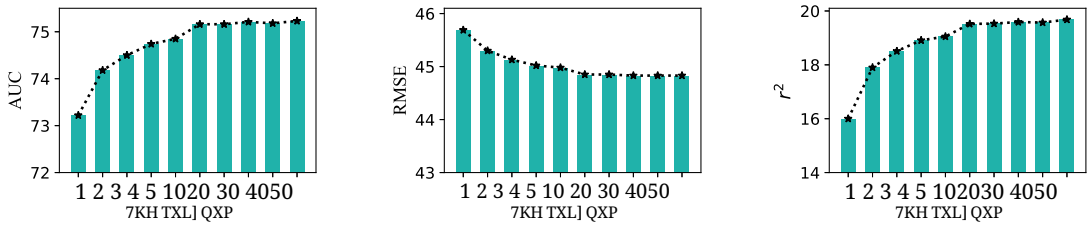


图4. 在 Eedi2020 上使用不同测验数量时 QKT 的性能（×100）。

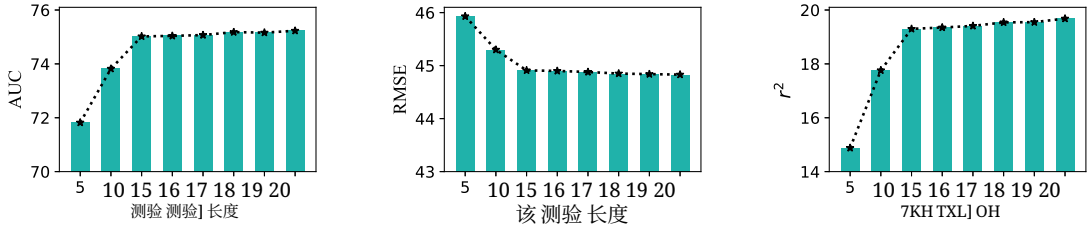


图5. 在 Eedi2020 上使用不同测验长度时 QKT 的性能（×100）。

注意编码器，并假设仅存在测验间长期知识替代，即学生的知识状态完全由 $\boldsymbol{sub}_j$ 表示，见不含 SUB 的 QKT。

- **不使用近期性感知注意力机制的 QKT**指不使用近期性感知注意力机制的 QKT，即移除了公式（9）中的近期性感知项 β_j^1, 和 β_j^2, 。

从表III所示结果中，我们可以得出一些有趣的结论。首先，移除知识替代建模会导致 QKT 的性能显著下降，这表明对学生过去和近期测验中具有相同或相似知识概念的测验一视同仁，会严重损害知识状态建模，因为学生在过去测验中的交互行为，在其完成后续测验后已不再可靠。其次，测验间的知识互补也很重要。如果我们只对测验间的知识替代进行建模，那么不含重叠知识概念的测验中所包含的信息就会丢失，QKT 的性能也如预期般下降。第三，衡量一次测验中各交互之间的相邻影响是必要的，这对测验内短期知识影响建模至关重要。最后，所提出的近期性感知注意力机制有助于更好地衡量知识替代，这验证了我们的假设：近期测验的影响更大。

G. 测验数量的影响（*RQ3*）

由于本文关注基于测验的知识追踪，本节将进一步评估每名学生的测验数量（即测验数量）如何影响 QKT 的表现。具体而言，我们比较了 QKT 在 Eedi2020 上针对 10 种不同测验数量的表现，分别为 1、2、3、4、5、10、20、30、40 和 50。相应结果见图 4。我们可以直接观察到，当测验数量较少时，QKT 的表现与测验数量呈正相关。此外，仔细观察图中数据还能发现更多有趣的现象。首先，当测验数量达到某个阈值时，QKT 的表现趋于稳定；从图 4 可以看出，该阈值约为 10。这一现象反映了测验数量的某种边际效应，启发我们设计更有效的测验组合，而不是要求学生完成更多测验，以便更好地估计其知识状态。其次，观察图 4 各子图中最左侧的柱形可以发现，如果只对一次测验进行建模，QKT 的表现较差。然而，纳入不止一次测验后，表现会大幅提升。这一现象表明，QKT 在对测验间的长期知识整合进行建模时能显著受益。

H. 测验长度的影响（*RQ4*）

与测验数量相对，我们还评估了QKT在不同测验长度（即每个测验中的交互长度）下的不同表现。具体而言，我们在Eedi2020上比较了QKT在8种不同测验长度下的表现，即5、10、15、16、17、18、19、20。相应结果如图5所示，从中可以发现，测验中交互越多，表现越好，这符合预期。原因是更多交互包含了关于学生与测验相关知识概念的知识状态的更可靠信息。与测验数量类似，我们可以发现测验长度存在一定的边际效应，即当测验长度达到某个阈值（图5中约为15）后，QKT的表现增长非常缓慢，这可以指导我们设计更有效的测验，用更少的习题来提高学生的学习效率。

VII. 结论与未来工作

本文从基于测验的视角研究知识追踪任务中的学生交互序列，并提出一种新颖的基于测验的知识追踪（QKT）方法，用于对基于测验的组织方式中的交互序列进行建模。我们首先分析并总结了学生测验交互序列的特点：在单次测验中，序列在较短时间内连续发生，而在不同测验之间，则以一定的间隔离散出现。

随后，基于上述特点，我们分别考虑了测验内短期知识影响，以及测验间长期知识替代和互补性，并在 QKT 中进一步设计了相应模块，对这些因素进行度量和整合，以追踪学生动态的知识状态。最后，我们在三个公开真实世界数据集上进行了大量实验，以评估 QKT 的有效性。结果表明，QKT 的表现优于现有最佳方法。对测验数量和测验长度影响的进一步分析显示，QKT 有望有益于测验设计。未来，我们将进一步探索利用时间信息（如答题时间和间隔时间），更精确地建模测验内和测验间关系。此外，我们还将尝试引入不同测验之间预定义的知识关系，以显式度量测验间的知识整合。

参考文献

[1]Pythagoras Karampiperis 和 Demetrios Sampson. 智能学习系统中的自适应学习对象选择. 交互式学习研究期刊, 15 (4):389–407, 2004.[2]Reid Bates 和 Samer Khasawneh. 自我效能与大学生对在线学习系统的感知和使用. 《计算机在人类行为中的应用》, 23(1):175–191, 2007.[3]John K Tarus、Zhendong Niu 和 Ghulam Mustafa. 基于知识的推荐：电子学习中基于本体的推荐系统综述.人工智能评论, 50(1):21–48, 2018.[4]Qi Liu、Shiwei Tong、Chuanren Liu、Hongke Zhao、Enhong Chen、Haiping Ma 和 Shijin Wang. 利用认知结构进行自适应学习. 在 第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集, 第627–635页, 2019.[5] Albert T Corbett 和 John R Anderson. 知识追踪：程序性知识获取的建模.*UMUAI*, 4(4):253–278, 1994.[6] Qi Liu、Shuanghong Shen、Zhenya Huang、Enhong Chen 和 Yonghe Zheng. 知识追踪综述. *arXivpreprintarXiv:2105.15106*, 2021.[7] Sepp Hochreiter 和 J¨urgen Schmidhuber. 长短期记忆. 神经计算, 9(8):1735–1780, 1997.[8] Chris Piech、Jonathan Bassen、Jonathan Huang、Surya Ganguli、Mehran Sahami、Leonidas J Guibas 和 Jascha Sohl-Dickstein. 深度知识追踪. 在 *NeurIPS*, 第505–513页, 2015.[9] Zachary A Pardos 和 Neil T Heffernan. 在知识追踪的贝叶斯网络实现中建模个性化. 在 *UMAP*, 第255–266页. Springer, 2010.[10]Michael V Yudelson、Kenneth R Koedinger 和 Geoffrey J Gordon. 个性化贝叶斯知识追踪模型. 在 *AIED*, 第171–180页. Springer, 2013.

[11]Shuanghong Shen, Qi Liu, Enhong Chen, Han Wu, Zhenya Huang, Weihao Zhao, Yu Su, Haiping Ma, and Shijin Wang. 卷积式知识追踪：对学生学习过程中个体差异的建模。*SIGIR '20*：第 *43* 届 *ACM* 信息检索研究与发展国际会议（虚拟会议，中国，*2020* 年 *7* 月），第 1857–1860 页，2020 年。[12]Yutao Wang and Neil T Heffernan. 将首次作答时间纳入知识追踪模型。国际教育数据挖掘学会, 2012 年。[13]Jos´e Gonz´alez-Brenes, Yun Huang, and Peter Brusilovsky. 知识追踪中的通用特征：对多项子技能、时间项目反应理论和专家知识进行建模。载于第七届教育数据挖掘国际会议，第 84–91 页。匹兹堡大学，2014 年。[14] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Mat-suo. 基于图的知识追踪：使用图神经网络对学生熟练程度进行建模。载于*2019* 年 *IEEE/WIC/ACM* 国际网络智能会议（*WI*），第 156–163 页。IEEE，2019 年。[15] Shi wei Tong, Qi Liu, Wei Huang, Zhenya Huang, En-hong Chen, Chuanren Liu, Haiping Ma, and Shijin Wang. 基于结构的知识追踪：一种影响传播视角。载于第 *19* 届 *IEEE* 国际数据挖掘会议论文集，2020 年。[16]Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 用于知识追踪的动态键值记忆网络。载于第 *26* 届万维网国际会议论文集，第 765–774 页，2017 年。[17]Shalini Pandey and George Karypis. 一种用于知识追踪的自注意力模型。*arXiv* 预印本 *arXiv:1907.06837*，2019 年。[18] Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. 情境感知的注意力知识追踪。KDD '20，第 2330–2339 页，美国纽约州纽约市，2020 年。[19] Libor Juhaˇn´ak, Jiˇr´ı Zounek, and Lucie Rohl´ıkov´a. 使用过程挖掘分析学习管理系统中学生的测验作答行为模式。《计算机在人类行为中的应用》，92：496–506，2019 年。[20] Emily Jensen, Tetsumichi Umada, Nicholas C Hunkins, Stephen Hutt, A Corinne Huggins-Manley, and Sidney K D' Mello. 你的行为预示着你的表现：利用在线学习环境中的活动特征前瞻性地建模学生测验表现。载于*LAK21*：第 *11* 届国际学习分析与知识会议，第 121–131 页，2021 年。[21] Ingrid AE Spanjers, Karen D K¨onings, Jimmie Leppink, Dani¨elle ML Verstegen, Nynke de Jong, Katarzyna Cza-b anowska, and Jeroen JG van Merrienboer. 混合式学习的理想境地：测验作为调节因素。教育研究评论，15：59–74，2015 年。[22]Zichao Wang, Angus Lamb, Evgeny Saveliev, Pash-mina Cameron, Yordan Zaykov, Jos´e Miguel Hern´andez-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 诊断性问题：NeurIPS 2020 教育挑战。*arXiv* 预印本 *arXiv:2007.12061*,

2020 年。[23] Gabriele Bard and Yana Weinstein. 问题顺序对测试表现评估的影响：偏差能否消除？《实验心理学季刊》，70(10)：2130–2140，2017 年。[24]Kyunghyun Cho, Bart Van Merri¨enboer, Caglar Gul-cehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 使用 RNN 编码器-解码器学习用于统计机器翻译的短语表示。*ArXiv:1406.1078*，2014 年。[25] Chen Lin and Min Chi. 干预式贝叶斯知识追踪：将教学干预纳入贝叶斯知识追踪。载于国际智能辅导系统会议，第 208–218 页。Springer，2016 年。[26] Prema Nedungadi and MS Remya. 在个性化聚类贝叶斯知识追踪（PC-BKT）模型中纳入遗忘因素。载于*2015* 年国际认知计算与信息处理会议（*CCIP*），第 1–5 页。IEEE，2015 年。[27] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 图神经网络综合综述。*IEEE* 神经网络与学习系统汇刊，32(1)：4–24，2021 年 1 月。[28]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 注意力就是一切。载于神经信息处理系统进展，第 5998–6008 页，2017 年。[29] Moyu Zhang, Xinning Zhu, Chunhong Zhang, Yang Ji, Feng Pan, and Changchuan Yin. 多因素感知双重注意力知识追踪。载于第 *30* 届 *ACM* 国际信息与知识管理会议论文集，第 2588–2597 页，美国纽约州纽约市，2021 年。计算机协会。[30]Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 练习感知知识追踪：面向学生表现预测的练习感知知识追踪。*IEEE* 知识与数据工程汇刊，33(1)：100–115，2019 年。[31] Shalini Pandey and Jaideep Srivastava. RKT：用于知识追踪的关系感知自注意力。CIKM '20，第 1205–1214 页，美国纽约州纽约市，2020 年。[32] Yunfei Liu, Yang Yang, Xianyu Chen, Jian Shen, Haifeng Zhang, and Yong Yu. 通过预训练问题嵌入改进知识追踪。第 1577–1583 页。国际人工智能联合会议组织，2020 年 7 月。[33] Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. 通过探究问题难度效应评估学生的动态知识状态。载于第 *45* 届 *ACM* 信息检索研究与发展国际会议论文集，第 427–437 页，2022 年。[34] Chenyang Wang, Weizhi Ma, Min Zhang, Chuancheng Lv, Fengyuan Wan, Huijie Lin, Taoran Tang, Yiqun Liu, and Shaoping Ma. 知识追踪中的时间交叉效应。国际网络搜索会议与数据挖掘》，第517–525页，美国纽约州纽约市，2021年。计算机协会。[35] Shuanghong Shen、Qi Liu、Enhong Chen、Zhenya Huang、Wei Huang、Yu Yin、Yu Su 和 Shijin Wang。学习过程一致的知识追踪。载于 第*27*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集，第1452–1460页，2021年。[36] Ting Long、Yunfei Liu、Jian Shen、Weinan Zhang 和 Yong Yu。结合个体认知与习得估计追踪知识状态。载于 第*44*届 *ACM* 信息检索研究与发展国际会议论文集，第173–182页，2021年。[37] Dongmin Shin、Yugeun Shim、Hangyeol Yu、Seewoo Lee、Byungsoo Kim 和 Youngduck Choi。Saint+：整合时间特征进行 EdNet 正确性预测。载于 *LAK21*：第*11*届国际学习分析与知识会议，第490–496页，2021年。[38] Fucai Ke、Weiqing Wang、Weicong Tan、Lan Du、Yuan Jin、Yujin Huang 和 Hongzhi Yin。HITSKT：用于会话感知知识追踪的分层 Transformer 模型。*arXiv* 预印本 *arXiv:2212.12139*，2022年。[39] Qi Liu。迈向新一代认知诊断。载于 第三十届国际人工智能联合会议论文集。国际人工智能联合会议组织，2021年8月。[40]Ronald K Hambleton、Hariharan Swaminathan 和 H Jane Rogers。项目反应理论基础。Sage，1991年。[41] Mark D Reckase。多维项目反应理论的过去与未来。应用心理测量，21(1)：25–36，1997年。[42] Mark D. Reckase。18 多维项目反应理论。载于 统计学手册，第607–642页。Elsevier，2006年。[43] Song Cheng、Qi Liu、Enhong Chen、Zai Huang、Zhenya Huang、Yiying Chen、Haiping Ma 和 Guoping Hu。DIRT：用于认知诊断的深度学习增强项目反应理论。载于 第*28*届 *ACM* 国际信息与知识管理会议论文集。ACM，2019年11月。[44] C. K. Yeung。Deep-IRT：使用项目反应理论使基于深度学习的知识追踪具有可解释性。*arXiv* 电子预印本，2019年。[45] Fei Wang、Qi Liu、Enhong Chen、Zhenya Huang、Yuying Chen、Yu Yin、Zai Huang 和 Shijin Wang。面向智能教育系统的神经认知诊断。载于 *AAAI*人工智能会议论文集，第34卷，第6153–6161页，2020年。[46]Yuqiang Zhou、Qi Liu、Jinze Wu、Fei Wang、Zhenya Huang、Wei Tong、Hui Xiong、Enhong Chen 和 Jianhui Ma。为学生学习中的认知诊断建模上下文感知特征。载于 第*27*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集。ACM，2021年8月。[47] Weibo Gao、Qi Liu、Zhenya Huang、Yu Yin、Haoyang Bi、Mu-Chun Wang、Jianhui Ma、Shijin Wang 和 Yu Su。

RCD：由关系图驱动的智能教育系统认知诊断。载于第 *44* 届 *ACM* 信息检索研究与发展国际会议论文集。ACM，2021年7月。[48] Mingyu Feng、Neil Heffernan 和 Kenneth Koedinger。通过在线辅导与评估系统应对评估挑战。*USER-ADAP*，19(3)：243–266，2009年。[49]Stephen H Edwards 和 Krishnan Panamalai Murali。Codeworkout：带有内置数据收集功能的短编程习题。载于 *2017*年 *ACM* 计算机科学教育创新与技术会议论文集，第188–193页，2017年。[50] E Joseph。参与度追踪：使用响应时间建模学生的 disengagement。人工智能教育：通过智能且具有社会信息意识的技术支持学习，125：88，2005年。[51] Neal M. Kingston 和 Neil J. Dorans。项目位置效应及其对项目反应理论等值和自适应测验的影响。应用心理测量，8(2)：147–154，1984年。doi：10.1177/014662168400800202。[52]Murray Glanzer 和 Anita R. Cunitz。自由回忆中的两种存储机制。言语学习与言语行为杂志，5(4)：351–360，1966年8月。[53] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。[54] Xavier Glorot 和 Yoshua Bengio。理解训练深度前馈神经网络的困难。载于 *AISTATS*，第249–256页，2010年。