·计算建模与人工智能·

# 过程还是结果？基于内部奖赏的篮球运动员情境特异性决策建模*

杨 柳[1] 章曦元[2] 周成林[1] 桑 标[**3]
（[1]上海体育大学心理学院，上海，200438）（[2]北京体育大学中国篮球运动学院，北京，100084）
（[3]上海市教育科学研究院，教育部教育大数据与教育决策实验室，上海，200032）

**摘 要** 篮球运动员在比赛中的决策往往是基于具体情境与自身经验的连续过程，这与强化学习理论的基本观点相符。然而，传统强化学习模型在捕捉多选项连续决策的动态特征方面存在不足。为此，本研究共招募 56 人（29 名篮球运动员），采用两阶段任务，并纳入改良后的内部增强模型，探讨情境与运动经验对决策过程的影响，并将其与其他经典模型进行比较。结果显示，内部增强模型在所有实验条件下均表现出最佳拟合效果。虽然篮球运动员和新手的决策均主要以结果为导向，但篮球运动员对过程目标的敏感性显著更高；此外，在篮球战术图情境下因信息复杂度高，个体更倾向于采用探索策略，导致决策时间延长。

**关键词** 强化学习 内部增强模型 篮球 决策 两阶段任务

## 1 引言

当篮球运动员在场上竞赛时，他们的决策往往是一个依据所处具体情境与自身经验相结合的连续过程（Rösch et al., 2021; van Maarseveen et al., 2018）。已有比赛数据表明，若球员在上一个进攻回合中成功命中三分球，则在下一回合中更倾向于再次选择三分出手；反之，若未命中三分球，则更倾向于避免再次尝试（Neiman & Loewenstein, 2011）。这表明篮球运动员会基于奖赏反馈调整后续的战术决策，这一特征与基于马尔可夫决策链（markov decision processes，MDP）的强化学习（reinforcement learning，RL）理论框架不谋而合：RL 模型将决策过程视为智能体（agent）与环境（environment）之间的交互，智能体通过在环境中执行一系列动作（actions），触发环境状态（states）的转移，并接收奖励信号（rewards）。通过这种连续的循环交互，智能体逐渐学习到能够最大化累积奖赏的策略（policy）（Sutton & Barto, 2018）。在 RL 框架中，决策策略通常分为无模型（model-free, MF）和基于模型（model-based, MB）两类（Shteingart & Loewenstein, 2014）。无模型策略基于过去的奖赏反馈，通过对历史行为进行直接强化来指导决策，无需了解环境的全貌（Shteingart et al., 2013）；而基于模型的策略则依赖于对环境结构的内在建模，通过推演未来可能的状态变化进行规划性决策（Huys et al., 2012）。篮球进攻战术的决策是一种高度依赖情境与经验的动态过程，其复杂性与独特性为 RL 理论在运动领域的验证和应用提供了重要研究契机。

两阶段任务（two-stage task）是强化学习 RL 框架中常用的范式，用于区分 MF 和基于 MB 决策策略（Daw et al., 2011）。该任务将决策过程分为两个连续阶段：在第一阶段，需要在两个选项间进行二元决策，每个选项以一定概率转移到第二阶段的某个状态。状态转移可分为常见转移（高概率，例如 70%）和罕见转移（低概率，例如 30%）。在第二阶段，参与者需要在两个选项中再次做出二元决策，这一决策可能得到奖赏，也可能没有奖赏，每个二阶段选项获得奖赏的概率会随着时间的推移而变化（图 1）。MF 策略的核心在于重复过去获得奖

* 本研究得到国家自然科学基金面上项目（32371129）的资助。
** 通讯作者：桑标，E-mail：bsang@psy.ecnu.edu.cn
DOI:10.16719/j.cnki.1671-6981.20260102

励的行为，而不考虑状态转移的具体类型；MB 策略则需要更全面地考量任务结构，通常表现为单试次逻辑回归（one-trial-back logistic regression）分析下，更高的状态转移系数和奖赏 × 状态转移的交互效应（Daw et al., 2011; Feher da Silva & Hare, 2018; Kevin et al., 2016; Luna et al., 2023）。对于篮球运动员来说，特定领域的情境能够促进他们的 MB 决策。与抽象条件相比（即刺激材料为难以理解的抽象符号），经验能使他们在篮球战术 图刺激条件下更有效地利用状态转移，并将状态转移与奖赏更加紧密地联系起来。随着研究人员对 RL 决策行为领域的深入探索，研究发现，人类的决策不仅局限于单一的 MF 或 MB 策略，还包括二者的混合策略及其他多种模型的策略，例如 the Unlucky-Symbol Algorithm 或 the Actor-Critic Algorithm （Feher da Silva & Hare, 2020）。这些模型无法通过标准的 MF-MB 框架进行解释（Feher da Silva & Hare, 2020）。尽管这些替代模型在两阶段任务中的表现看似“不正确”，但它们可能同样反映了有效的决策过程，甚至在特定情境下更能解释智能体的行为。因此，在两阶段任务中“纯粹的”MB 策略的缺失不应仅仅归因于参与者对任务结构的错误理解，也许其他潜在模型可能更适合解释该决策的过程。

传统 MF-MB 框架主要通过外部奖赏进行价值更新，但根据目标设定理论（goal setting theory），在运动情境中，目标可设置为过程目标（process goal）和结果目标（outcome goal）：前者关注战术执行等具体行为，后者则聚焦于得分或胜利等外部结 果 （Kingston & Hardy, 1997; Locke, 1968; Mullen & Hardy, 2010）。结合篮球项目的特点，过程目标关注战术的执行情况，例如队友间的配合、跑位是否到位、防守是否有效等；而结果目标则直接指向得分或胜利等可量化的外部奖赏。在许多情况下，运动员可能成功实现了过程目标，但未能达到结果目标。例如，在一次进攻回合中，球员成功完成了预定的战术执行（如掩护、传球或跑位），但最终投篮未命中。该情况下，虽然结果不理想，但从过程目标的角度看，战术执行的成功本身就带来了内部奖赏。这一核心思想与 Molinaro and Collins（2023）提出了一种名为“内部增强”（intrinsically enhanced）的计算模型不谋而合。该模型假设智能体在决策时会将获得奖赏的外部奖赏信号与目标达成（goal achievement）的内部奖赏信号结合起来，并对二者赋予不同的权重，从而实现情境特异性的价值评估。这一观点在传统 MF-MB 框架中尚未被充分考虑。然而，该模型只有学习率 α 和内部奖赏权重 ω 两个参数，其参数设定较为简单；并且模型未采用 soft-max 函数来将各选项的价值转化为选择概率，限制了其对于决策不确定性的刻画能力（Sutton & Barto, 2018）。因此，基于内部增强模型在连续决策建模中的局限性，本研究结合篮球运动的情境需求，对其结构进行了针对性改进，以更准确地捕捉多选项决策中内部与外部奖赏整合的动态特征。

综上所述，篮球运动员在连续决策过程中，如何将“过程目标”所带来的内部奖赏与“得分或胜利”所带来的外部奖赏相结合，是本研究首要探讨的科学问题。此外，若在现有 RL 理论基础上引入并改进内部增强模型，能否更好地解释与预测运动员的决策行为，则是本研究的另一关注重点。基于以上两个科学问题，本研究在 Molinaro 与 Collins（2023）的工作基础上，对两阶段任务和内部增强模型进行了改进，以更好地适应探测篮球运动员在特定情境下决策机制的需求。首先，在任务设计方面，为弥补传统任务生态效度较低的问题，我们将一阶段战术图保持镜像对称，二阶段则设计为更贴近实际比赛的决策选项（例如分球给顺下队友或直接投篮）。此外，为避免战术意义引发的反应偏好与奖赏概率波动之间的混淆，本研究将所有二阶段选项的奖赏概率统一设置为 50%。在计算模型方面，本研究以内部增强模型为基础，结合传统 RL 模型的特点，将状态转移的成功与否作为内部增强的信号，并引入 soft-max 函数，用以将价值转化为选择概率。最后，对每个阶段选项进行独立的价值更新，从而更全面地模拟个体的动态决策过程。

## 2　方法

### 2.1　实验设计

本研究采用 2（实验材料：抽象符号、篮球战术图）× 2（组别：无经验组、篮球运动员）混合实验设计，其中实验材料为被试内变量，组别为被试间变量。

### 2.2　被试

样本量的设定参考了 Brysbaert （2019） 的建议，并通过 G*power 3.1 进行估算 （Cunningham & McCrum-Gardner, 2007），选择重复测量方差分析，

参数为 $\alpha$ =.05，Power（1-$\beta$ ）= .80，Effect size $f$ =.25，得出需要 34 名被试。本实验共招募 56 人（其中 28 男），平均年龄 22.55 ± 2.84 岁。其中无经验组 27 人（其中 10 男），平均年龄 22.90 ± 3.56 岁，均未有过任何篮球运动和观赛经历。运动员组 29 人（其中 18 男），均为国家二级及以上篮球运动员，平均年龄 22.20 ± 1.95 岁，平均篮球运动年限 9.48 ± 2.52 年，平均每周训练 7.69 ± 3.46 小时。事后统计功效分析表明，实际统计效力 Power （1-$\beta$ ） = .96。所有被试自愿参与实验，均阅读知情同意书并签字，实验结束后发放礼品作为报酬。实验已通过伦理审查。

### 2.3 实验材料

两阶段任务沿用 Daw 等人（2011） 的原始任务的框架，使用 MATLAB R2021b 软件开发，结合 Psychtoolbox（版本 3.0.19）进行呈现。试次开始时，屏幕显示一个注视点，持续 1 秒。注视点消失后，进入一阶段（S1）。被试需要在该阶段呈现的两个选项中做出初始决策，随后屏幕显示 1 秒的黑屏，并进入二阶段（S2）。二阶段呈现的刺激由一阶段的决策结果决定。例如，选择选项 1 后，有 70% 的概率进入 S2-1（常见转移）和 30% 的概率进入 S2-2（罕见转移）。在第二阶段，参与者需要再次做出决策，并收到反馈（奖励或无奖励）。每个选项的奖励概率为 50%（图 1）。为了避免刺激 - 反应联结的干扰，任务中的刺激位置以随机方式呈现（Luna et al., 2023; Luque et al., 2020）。整个任务包含两个组块：抽象符号情境（图 1A）和篮球战术图情境（图 2），每个组块包含 150 个试次。

抽象符号情境下刺激与指导语与原始任务保持一致 （Daw et al., 2011）。实验中所使用的刺激材料为藏文字母，该类符号在后续研究中常被作为“抽象条件”加以沿用 （Campbell et al., 2025; Feher da Silva & Hare, 2020; Feher da Silva et al., 2023; Luna et al., 2023）。本研究所选取的藏文字母包括“ཀ、ཁ、ཅ、ཆ、ཉ、ཏ”。所有被试均非藏族，且无藏语学习经历，因此对上述符号不具备语义或语音层面的熟悉性。

篮球战术图情境下的一阶段的图片代表挡拆战术的发起阶段，以完全镜像方式呈现，二阶段战术图代表战术完成阶段，战术情境一致但选项不同，可选择传球给挡拆顺下的队友或三分球投篮。S2-1 与 S2-2 以完全镜像方式呈现。指导语修改为符合篮球战术执行的描述：“这是一个决策实验，在第一阶段，代表两个战术的发起阶段，按 ← 选择左边的战术，按 → 选择右边的战术。每个战术通常会以 70% 的概率执行成功，至预想状态；或者以 30% 的概率到执行失败，至非预想状态。在第二阶段，代

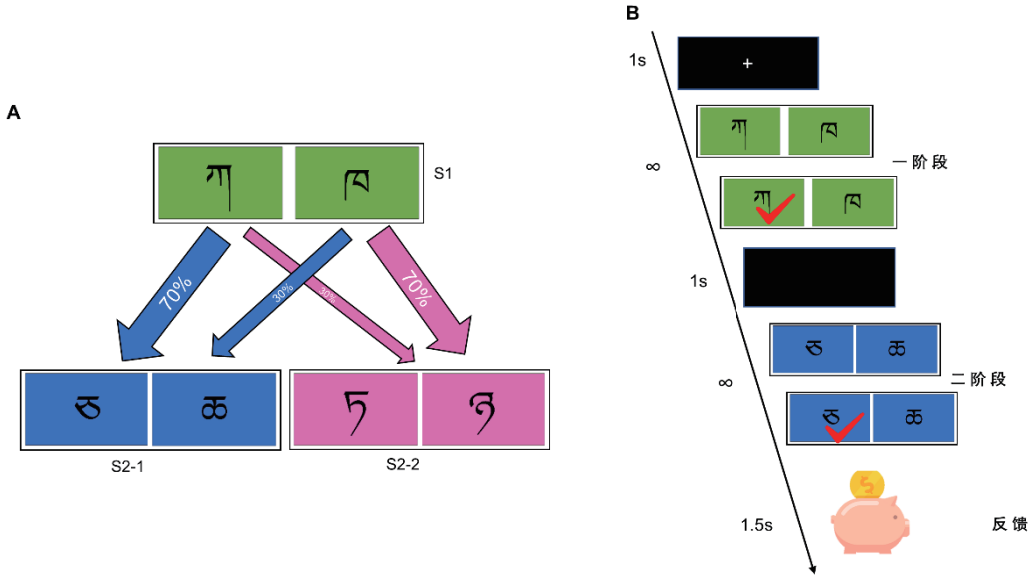


**图 1 两阶段任务流程图**

注：(A) 状态转移结构图。参与者在第一阶段（S1，绿色框）选择其中一个符号后，将以概率性的方式转移至第二阶段（S2–1 或 S2–2）。每个第一阶段的选项与第二阶段的状态有固定的转移概率（如图中箭头所示，70% 与 30%）。第二阶段包括两组状态（S2–1 为蓝色框，S2–2 为粉色框），每组中包含两个符号可供选择。(B) 试次流程图。每一试次开始于一个持续 1 秒的注视点（+），随后呈现第一阶段的两个选项，被试做出选择后（红色勾选标记所示），经过 1 秒空屏期进入第二阶段。被试在第二阶段中选择一个选项，最终获得反馈。

表战术的终结，按 ← 代表选择左边的选项，按 → 代表选择右边的选项，这个决策可能会成功，也可能失败。你需要不断地探索和学习以找出最佳的决策。”因此，若被试选择左侧挡拆并走左路的进攻战术，则战术成功的标志是二阶段所呈现的选项中，持球人位于左侧区域（概率为 70%）；若未成功转移至该状态，则表示战术执行失败（细节如图 2 所示）。

### 2.4　实验程序

被试阅读并签署《知情同意书》后，通过抽取随机数决定第一个组块的实验条件。在每个组块开始前，进行 15 个试次的任务练习，随后进入正式实验。实验结束后，向被试表达感谢并发放礼物。

### 2.5　单试次逻辑回归

该方法为分层逻辑回归模型（hierarchical logistic regression model），其参数通过贝叶斯计算方法估计。纳入模型的基础观测数据为：当前试次中，被试在一阶段的决策行为（记作 1 或 2）；前一个试次中的状态转移类型、是否获得奖赏。

被预测变量为$p_{stay}$（保持前一个试次一阶段选项的概率），预测变量为$x_r$（前一试次是否获得奖赏：若获得奖赏，则为 +1；没有则为 -1）；$x_r$（前一试次的转移类型：常见转移则为 +1；罕见转移则为 -1）；及二者的交互作用（公式 1）。

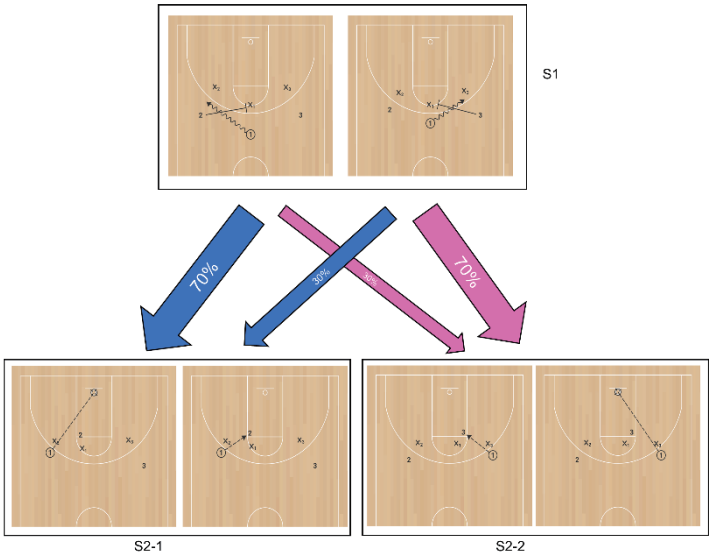


**图 2　篮球战术图情境两阶段任务示意图**

注：S1 阶段呈现两种不同的战术开局选择：左侧为 1 号球员从左侧发起突破，右侧为 1 号球员从右侧发起突破。在每个试次中，被试需从这两个选项中做出选择。每个战术开局将以概率性的方式转移至第二阶段的两个战术状态之一：S2–1 或 S2–2。左侧战术开局有 70% 的概率转移至 S2–1，30% 概率转移至 S2–2；右侧开局则相反。S2 阶段分别呈现两个不同的战术选择图：S2–1 左侧战术面对防守人直接投篮；S2–1 右侧战术为分球给篮下空位的队友。S2–2 战术与 S2–1 战术镜像对称。

$$p_{stay} = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_r x_r + \beta_t x_r + \beta_{r\times t} x_r x_t)\right]}$$ 公式 1

在每对试次中，若被试在后一试次中选择与前一试次相同的第一阶段行为，则变量 $y$（$p_{stay}$）取值为 1；若选择不同，则变量 $y$ 取值为 0。模型参数的先验分布与 Feher da Silva 等人 (2023) 保持一致。模型使用 Stan 语言进行编码，并在 Python （版本 3.12）环境下通过 PyStan（版本 3.10） 软件包生成 64000 份样本（来自 4 条链，每条链包含 32000 份样本，前 16000 份用于热身）。所有参数的 $\hat{R}$ ≈ 1.0 表明链已收敛。

### 2.6　计算模型

所有模型的编码语言、计算环境以及样本生成参数均与单试次逻辑回归模型一致。

#### 2.6.1　无模型（**Model-Free**，**MF**）算法

MF 算法为基于时间差分学习的 SARSA ($\lambda$)（state-action-reward-state-action with eligibility trace, $\lambda$）算法（Daw et al., 2011; Rummery & Niranjan, 1994）。纳入模型的基础观测数据为：当前试次中，被试在

一阶段的决策行为（记作 1 或 2）、二阶段的决策行为（记作 1 至 4）以及是否获得奖赏（若获得奖赏，则为 1；没有则为 0）。

任务中每个试次 $t$ 的每个阶段 $i$ 记作$s_{i,t}$，发生的决策行为记作$a_{i,t}$，$\alpha_i$为两个任务阶段中分别独立的学习率。每个试次的价值预期值$Q_{MF}$根据学习率 $\alpha$ 和奖赏预测误差$\delta$（reward prediction error, RPE）更新，其规则如公式 2 所示：

$$Q_{MF}(s_{i,t},a_{i,t}) = Q_{MF}(s_{i,t},a_{i,t}) + \alpha_i\delta_{i,t} \quad \text{公式 2}$$

其中，

$$\delta_{i,t} = r_{i,t}+Q_{MF}(s_{i+1,t},a_{i+1,t}) - Q_{MF}(s_{i,t},a_{i,t}) \quad \text{公式 3}$$

需要注意的是，一阶段的决策本身不直接获得奖赏（即$r_{1,t}=0$），而是通过二阶段的价值预期值$Q_{MF}(s_{2,t},a_{2,t})$来计算 RPE。由于该任务没有第三阶段，因此$Q_{MF}(s_{3,t},a_{3,t})$被定义为 0，第二阶段的 RPE 由决定。最后，使用资格迹（eligibility trace）$\lambda$ 调节第二阶段的 RPE，将其用于更新第一阶段的价值预期值$Q_{MF}(s_{1,t},a_{1,t})$，从而实现跨阶段价值更新，其规则如公式 4 所示：

$$Q_{MF}(s_{1,t},a_{1,t}) = Q_{MF}(s_{1,t},a_{1,t}) + \alpha_1\lambda\delta_{2,t} \quad \text{公式 4}$$

#### 2.6.2 基于模型（**model-based**，**MB**）算法

MB 算法首先通过学习状态转移概率，构建一阶段决策对第二阶段状态的映射关系；随后基于第二阶段的即时奖励值，利用 Bellman 方程对“状态 - 行为”价值进行迭代计算，从而得出累积价值 (Daw et al., 2011)。纳入模型的基础观测数据与 MF 算法相同。

任务共包含三个状态（即，一阶段 $S_A$；两个二阶段 $S_B$ 和 $S_C$）和两种行为（即，$a_A$ 和 $a_B$），根据任务结构，转移概率如下：$P(S_B|S_A,a_A)$=0.7, $P(S_C|S_A,a_A)$=0.3; 或 $P(S_C|S_A,a_B)$= 0.7, $P(S_B|S_A,a_B)$=0.8。

随后，使用 Bellman 方程更新一阶段的价值预期值$Q_{MB}(s_A,a_j)$，其更新规则如公式 5 所示：

$$Q_{MB}(s_A,a_j) = P(s_B|s_A,a_j)\max_{a\in\{a_A,a_B\}} Q_{MF}(s_B,a) + P(s_C|s_A,a_j)\max_{a\in\{a_A,a_B\}} Q_{MF}(s_C,a) \quad \text{公式 5}$$

随后为引入$\omega$参数为$Q_{MB}$和$Q_{MF}$进行加权，控制二者的相对贡献，以此计算$Q_{net}(s_A,a_j)$（公式 6）。

$$Q_{net}(s_A,a_j) = \omega Q_{MB}(s_A,a_j) + (1-\omega)Q_{MF}(s_B,a_j) \quad \text{公式 6}$$

最后通过 soft-max 函数计算选择行为的概率（公式 7）：

$$P(a_{i,t}=a|s_{i,t}) = \frac{\exp(\beta_i[Q_{net}(s_{i,t},a)+\rho\cdot rep(a)])}{\sum a'\exp(\beta_i[Q_{net}(s_{i,t},a')+\rho\cdot rep(a')]} \quad \text{公式 7}$$

其中$\beta_i$为两个阶段的逆温（inverse temperature）参数，用于决定每个阶段的探索 - 利用率（exploration – exploitation rate）；$\rho$ 为坚持（perseveration）参数，表示智能体在下一试次中重复上一试次一阶段动作$rep(a)$的倾向； 为 1 时代表智能体重复执行了上一个试次中的一阶段动作$a$，反之为 0。综上，在 MB 算法中共有 7 个自由参数（$\beta_1,\beta_2,\alpha_1,\alpha_2,\omega,\lambda,\rho$），当$\omega$为 0 时，智能体采用纯 MF 策略，当$\omega$为 1 时，智能体则采用纯 MB 策略。

#### 2.6.3 内部增强（**intrinsically enhanced**）算法

该算法在原始内部增强模型的基础上进行改进，以适配两阶段任务，核心思想是将状态转移视为内部奖赏信号（$ir_t$），并结合二阶段的即时奖励，对一阶段两个选项（$k$）的价值预期值$Q_{k,t}$分别进行迭代更新 (Molinaro & Collins, 2023)。纳入模型的基础观测数据为：当前试次中，被试在一阶段的决策行为（记作 1 或 2）、状态转移类型（若转移结果为常见则记作 1；罕见转移记作 0）以及是否获得奖赏（若获得奖赏，则为 1；没有则为 0）。

算法首先将$Q_{k,t}$设为 0，并根据状态转移的结果判断。随后根据 Rescorla - Wagner 算法 (Sutton & Barto, 2018)，结合二阶段的即时奖赏进行迭代更新，具体规则如公式 8 所示：

$$Q_{k,t+1} = \theta ir_t + (1-\theta)[Q_{k,t} + \alpha_2(r_t - Q_{k,t})] \quad \text{公式 8}$$

其中$\alpha_2$为二阶段即时奖赏的学习率。$\theta$代表内部奖赏信号权重，当$\theta$为 1 时，智能体价值更新完全由内部奖赏驱动；当$\theta$为 0 时，则完全由外部奖赏驱动（更新方式退化为标准的 Rescorla - Wagner 形式）。

最后通过 soft-max 函数计算选择行为的概率（公式 9）：

$$P(a_k) = \frac{\exp(\beta Q_k)}{\sum k'\exp(\beta Q_{k'})} \quad \text{公式 9}$$

其中$\beta$为逆温参数，意义与 MB 模型相同。

综上，内部增强模型中共有$\alpha_2$、$\theta$、$\beta$三个核心参数，协同表征了个体在两阶段任务中整合内部状态评估与外部反馈学习的动态决策机制。

### 2.7 模型比较

在数据预处理阶段，首先计算每个被试在每个组块内反应时的 $z$ 分数，然后剔除 $|z| > 2$ 的异常试次，以减少极端值对数据分析的干扰。随后在 Python（版本 3.12）环境下，通过 ArviZ（版本 0.20）软件包进行模型比较。本研究采用 WAIC（widely

applicable information criterion）方法的核心指标——ELPD（expected log pointwise predictive density），对模型的拟合优度和预测能力进行评估 (Vehtari et al., 2017)。ELPD_WAIC 通过估算模型对被试在每一试次中决策行为的对数似然，并结合模型有效参数的数量（$\hat{p}_{waic}$），从而对过拟合进行惩罚（郭鸣谦等，2024）。通过比较 4 种条件下（新手—抽象、新手—篮球、专家—抽象、专家—篮球）各模型的 ELPD_WAIC 值，直接判断模型的拟合表现，其中值越高表示模型的拟合效果越好。

### 2.8 数据分析

经模型比较后，选择每个条件下的最优拟合模型，并对该模型的参数以及两个决策阶段的反应时进行 2（实验材料：抽象符号、篮球战术图）×2（组别：无经验组、篮球运动员）重复测量方差分析，其中实验材料为被试内变量，组别为被试间变量。当球形假设（sphericity assumption）不满足时，采用 Greenhouse-Geisser（GG）校正，效果量使用（general eta squared）进行测量。

## 3 结果

### 3.1 模型比较

模型比较结果表明，在 4 种实验条件下，内部增强模型的 ELPD_WAIC 值均高于其他模型（详见表 1）。因此，可以推断内部增强模型为最优拟合模型。故在后续的模型参数分析中，仅对内部增强模型生成的参数进行重复测量方差分析。

### 3.2 模型参数

对内部增强模型生成的 3 个参数分别进行 2（实验材料：抽象符号、篮球战术图）×2（组别：无经验组、篮球运动员）重复测量方差分析，结果发现：$\alpha_2$ 系数组别主效应不显著，$F(1, 54) = .71$，$p > .05$；实验材料主效应不显著，$F(1, 54) = .93$，$p > .05$；交互作用不显著，$F(1, 54) = .84$，$p > .05$（详见图 3A）。

$\theta$ 系数组别主效应显著，$F(1, 54) = 4.43$，$p < .05$，篮球运动员（.33 ± .20）高于无经验组（.24 ± .19）；实验材料主效应不显著，$F(1, 54) = .86$，$p > .05$；交互作用不显著，$F(1, 54) = .04$，$p > .05$（详见图 3B）。

表 1　不同实验条件下各模型拟合结果（ELPD_WAIC）

| 实验条件 | 模型 | | | |
|---|---|---|---|---|
| | 单试次逻辑回归 ($M \pm SE$) | MF 模型($M \pm SE$) | MB 模型($M \pm SE$) | 内部增强模型($M \pm SE$) |
| 新手—抽象 | -1853.81 ± 30.75 | -6645.34 ± 184.2 | -6405.02 ± 199.06 | -1743.42 ± 28.92 |
| 新手—篮球 | -2326.23 ± 23.37 | -7047.86 ± 152.03 | -6960.67 ± 160.83 | -2162.06 ± 26.49 |
| 专家—抽象 | -2318.96 ± 25.90 | -7671.37 ± 144.25 | -7494.64 ± 173.41 | -2236.83 ± 26.79 |
| 专家—篮球 | -2759.16 ± 17.17 | -7873 ± 126.51 | -7795.59 ± 133.25 | -2598.52 ± 21.29 |

注：MF 为 Model-Free 模型；MB 为 Model-Based 模型。ELPD_WAIC 值越高，模型拟合结果越好。

$\beta$ 系数组别主效应不显著，$F(1, 54) = 3.87$，$p > .05$；实验材料主效应显著，$F(1, 54) = 1.90$，$p < .01$，$\eta_p^2 = .058$，抽象符号情境（3.89 ± 3.97）显著高于篮球战术图情境（2.26 ± 2.82）；交互作用不显著，$F(1, 54) = .76$，$p > .05$（详见图 3C）。

### 3.3 反应时

对一阶段和二阶段得反应时分别进行 2（实验材料：抽象符号、篮球战术图）×2（组别：无经验组、篮球运动员）重复测量方差分析，结果发现：一阶段反应时组别主效应不显著，$F(1, 54) = .41$，$p > .05$；实验材料主效应显著，$F(1, 54) = 36.26$，$p < .001$，$\eta_p^2 = .222$，篮球战术图情境下（1283.27 ± 673.81）反应时显著高于抽象材料情境（741.26 ± 274.75）；交互作用不显著，$F(1, 54) = .16$，$p > .05$（详见图 4A）。

二阶段反应时组别主效应不显著，$F(1, 54) = .15$，$p > .05$；实验材料主效应显著，$F(1, 54) = 36.26$，$p < .001$，$\eta_p^2 = .085$，篮球战术图情境下（1289.81 ± 649.79）反应时显著高于抽象材料情境（96.54 ± 408.98）；交互作用不显著，$F(1, 54) = .70$，$p > .05$（详见图 4B）。

## 4 讨论

本研究通过 RL 框架与两阶段决策任务，结合认知计算建模技术，深入探讨了篮球运动员的决策特征。经过修改的内部增强模型在数据拟合方面优于传统的最优模型（即单试次逻辑回归模型），验证了该模型在揭示篮球运动员决策过程中的作用。

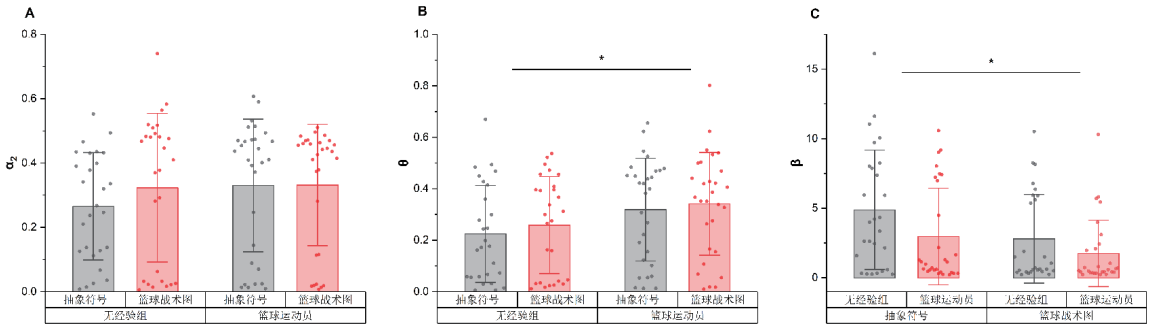


**图 3 内部增强模型参数重复测量方差分析结果**

注：(A) 学习率系数$\alpha_2$；(B) 内部奖赏信号权重系数$\theta$；(C) 逆温参数$\beta$。误差棒代表 ±1 标准差，散点代表单个被试数据点，* 代表 $p < .05$。

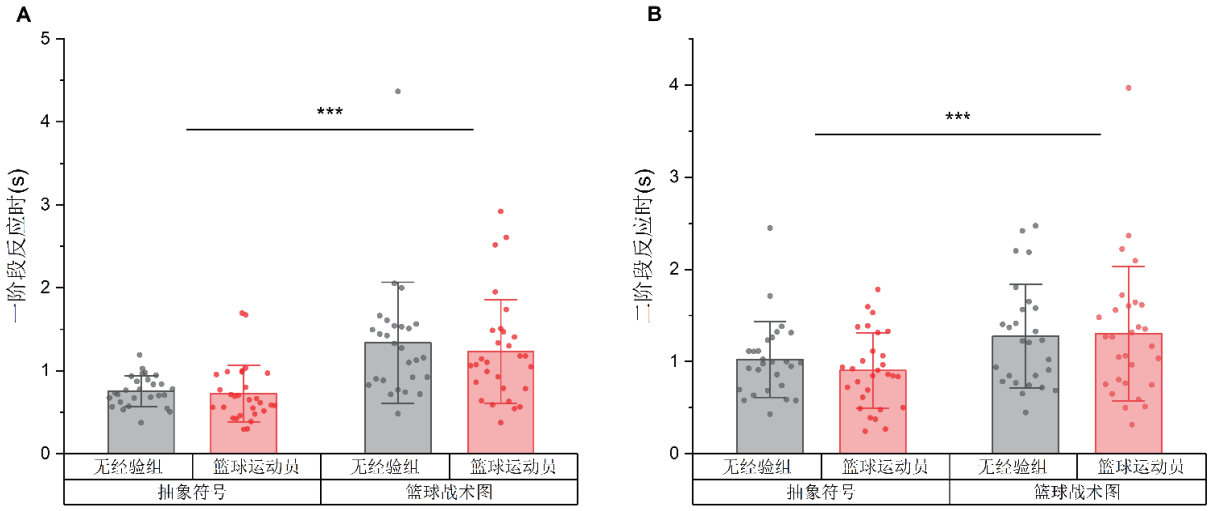


**图 4 反应时重复测量方差分析结果**

注：(A) 一阶段反应时；(B) 二阶段反应时。误差棒代表 ±1 标准差，散点代表单个被试数据点，*** 代表 $p < .001$。

已有大量证据表明，人类的价值判断具有显著的情境依赖性，这种现象在需要基于 RL 的决策任务中尤为明显（Bavard et al., 2021; Louie & De Martino, 2014; Palminteri et al., 2015; Rangel & Clithero, 2012）。然而，传统的 MF 和 MB 模型均高度依赖外部奖赏，难以对过程目标进行表达；单试次逻辑回归模型（目前研究阶段的最优拟合模型）仅基于前一个试次的状态转移与奖赏反馈预测当前决策，虽然能捕捉局部的反馈效应，但不具备长期策略更新和价值累计的能力。内部增强模型通过整合外部奖赏与内部奖赏信息，实现了对部分结果的重新标定，从而有效解释了个体在不同情境下对相同奖励进行差异化价值评估的现象，更好地模拟了现实中基于情境驱动的决策过程（Molinaro & Collins, 2023）。目前鲜有研究将内部增强模型纳入两阶段决策任务中，并同时考量经验和情境对个体决策行为的影响。人类决策行为可能并不完全符合传统两阶段任务中假设的 MF 与 MB 学习的二分法，而是往往表现为多种策略的综合运用，体现出更为复杂和灵活的决策机制（Feher da Silva & Hare, 2020）。模型比较的结果表明，经修改后的内部增强模型在所有实验条件下均呈现出最优拟合效果。这一优势主要源于该模型能够整合来自任务执行过程中的内部奖赏和任务执行结果的外部奖赏，并且能够灵活适应不同的情境、实时更新策略，以反映个体决策时的认知过程。这不仅验证了其在捕捉情境依赖性方面的理论优势，更充分反映了篮球运动经验如何影响个体在复杂情境下的动态决策特征。因此，篮球运动员在实际战术选择中可能不仅关注最终是否得分，也会基于战术执行过程中是否“顺利”这一内部反馈信号动态调整策略偏好，而这一过程特征正是内部增强模型所成功捕捉的。

$\alpha_2$ 系数的结果表明，当奖赏概率和绝对价值在各条件下保持一致时，决策过程中直接反馈的影响

表现出较高的稳定性，这可能源于一种基础且自动化的奖赏处理机制（Steffen et al., 2011）。这种机制在神经和计算层面上具有高度鲁棒性，不易受专业经验或刺激特征的调节（Levy & Glimcher, 2012）。价值作为连接决策行为与强化学习的核心，其候选行动的不同价值预期会影响决策行为，并通过反馈对行动进行持续修正，从而构成学习过程（O'Doherty et al., 2017）。尽管个体在高级策略选择上可能因经验和情境而存在差异，但直接反馈信号的初步处理保持一致，为整个决策过程提供了稳定基础。这种稳定性不仅确保了基本奖励信息的可靠编码，同时也使内部增强模型能够在稳定的低层反馈基础上，有效整合其他高层次调控因素的影响，从而更准确地反映实际决策行为的复杂性。

在决策过程中，篮球运动员相比新手，内部奖赏的权重更高，更易受到过程目标（如战术执行）带来的内部奖赏的影响。值得注意的是，内部增强模型结果显示，篮球运动员的 $\theta$ 参数约为 0.33，低于 0.5 的平衡点，这表明相比起无经验组，虽然篮球运动员在训练和比赛中对过程目标有更高的敏感性，但其决策过程仍主要受到外部奖赏（例如比赛结果、得分等）的驱动。这一结果符合目标设定理论：运动员在决策过程中往往不会单纯依赖于过程目标或结果目标，而是采用两者相结合的策略，并予以不同的权重（Jeong et al., 2021; Locke & Latham, 2013）。过程目标关注具体技术动作和战术执行，帮助运动员在竞技环境中迅速调整和优化表现；而结果目标则侧重于最终的得分或胜负，二者的混合不仅使运动员在决策时能够关注眼前的战术执行细节，还能始终保持对胜利的追求，是维持高水平运动表现的关键（Burton, 1989; Kingston & Hardy, 1997）。篮球运动员相比新手，其内部奖赏权重更高可能源于以下几点。首先，经过长期专业训练和比赛经验积累，篮球运动员对过程目标的认知和敏感性显著提高，使他们能够更准确地捕捉到与竞技表现相关的内部奖赏信号；新手由于经验不足，对这些细微的过程目标尚未形成稳定的内在评价体系，因而更依赖于外部结果目标（Ericsson et al., 1993; Locke, 2000）。其次，目标通过指引功能（directive function）影响运动表现，明确且具体的目标能够激发运动员的内部动机和自我调节能力。篮球运动员往往将过程目标内化为一种自我激励机制，从而在决策时给予更高的权重; 而新手缺乏这种内化过程，无法充分利用过程目标带来的正向反馈，进而更多地将价值更新的权重寄于外部的直接反馈（Locke & Latham, 2002）。最后，经验与目标对绩效的影响可能具有独立效应：当个体拥有与任务相关的既有经验（尽管这些经验未被直接测量）时，目标的设定能够激活这些经验；然而，较高难度的目标可能仅仅加速或强化这些知识的应用，而不会激活更多新的知识（Locke, 2000）。此外，$\theta$ 参数并未出现交互作用，这与 Song 等人（2025）关于视觉运动表征中自动化预测的元分析结果一致，即经验在任务中体现的专项优势具有普遍性，不仅体现在领域特定刺激的表征中，也适用于一般性刺激。鉴于运动预判与运动决策之间存在密切关联，因此这种因经验积累而形成的决策特点也很有可能在更广泛的情境中得到体现。

在 RL 框架中，利用（exploitation）策略指的是智能体根据先前的学习经验，选择当前看来最优的行动，从而降低决策的不确定性；而探索（exploration）策略则需要对不确定的备选方案进行评估，使得决策更具随机性。实验结果表明，相比起抽象刺激材料，篮球战术图刺激下参数 $\beta$ 更低，表明被试更倾向于采取探索策略，从而增加决策随机性。并且，需要注意的是，篮球情境同时显著延长了决策反应时。这与利用策略通常依赖自动化处理、反应迅速；而探索策略则需要信息获取和策略试探的观点相一致（Sang et al., 2020）。篮球战术图相比抽象材料展示的信息量更大且更为复杂，即使无经验者无法理解其含义，也会因高信息复杂度显著增加认知负荷。基于认知负荷理论，高信息复杂度要求个体投入更多认知资源来处理和整合信息，这种情境促使他们更倾向于探索策略，并延长反应时间（Sweller, 2011; Verschuere et al., 2018; Wojciechowski et al., 2025）。此外，对于一阶段的反应时，尽管组别主效应未达显著，但该边缘显著性可能反映了运动经验在该阶段反应速度上的初步差异趋势，这一趋势应在未来的研究中采用更大的样本量进一步检验。

综上，本研究使用内部增强模型揭示了运动经验和情境对决策过程的影响，为理解运动员的决策机制提供了量化指标。教练员或训练团队可以利用这些指标监测运动员在训练过程中的决策变化，并以此调整训练方案。例如，对新手或运动水平较低的运动员更多地引导其关注动作或战术执行的过

程，以建立基本的战术认知和动作标准，进一步强化内在奖励系统的作用；而对于经验丰富的运动员，则侧重于加强过程导向和结果导向的平衡，使运动员保持更为平稳和积极的心理状态，促使运动员在追求个人目标的同时关注团队整体目标，从而增强团队凝聚力，提高整体竞争力（Altfeld et al., 2017; Terner & Franks, 2021）。此外，训练方案也可考虑加入具有较高信息复杂度的模拟任务，迫使运动员在高认知负荷条件下快速识别关键信息，在提升对复杂视觉信息处理能力的同时，加速自动化决策的形成，从而缩短反应时间，提高赛场上的决策效率。

然而，本研究亦存在以下局限与不足：首先，尽管本研究对两阶段任务进行了优化和改良，但在提升生态学效度方面仍存在改进空间。未来的研究可以考虑采用更为真实的竞技情境，例如引入虚拟现实（VR）或动态战术图像，使实验环境更贴近真实比赛；也可考虑在强化学习框架下扩展更复杂的多选项战术模拟，以进一步提升对真实竞赛场景的解释力。其次，本研究仅二级运动员与新手之间的决策差异进行了比较，而未涵盖精英运动员与二级运动员之间的对比。未来的研究应扩大样本范围，纳入精英运动员，从而探究不同水平运动员在决策机制上的差异，为次精英向精英水平的跨越提供理论支持。最后，本研究仅对聚焦于篮球专项的决策特点，因此将研究结论推广到其他运动项目或领域时需谨慎。未来的研究应考虑涵盖不同类型的运动项目，以验证和比较各领域中决策机制的共性与差异，从而为内部增强模型的通用性提供理论依据。

## 5 结论

综上所述，本研究表明内部增强模型在解释和预测个体决策行为方面表现出更高的有效性。虽然篮球运动员和新手的决策均主要以结果为导向，但篮球运动员对过程目标的敏感性显著更高。此外，个体在面临复杂的信息时认知负荷更高，导致决策时间延长，并更倾向采用探索策略。

### 参考文献


郭鸣谦，潘晚坷，胡传鹏 . (2024). 认知建模中模型比较的方法 . *心理科学进展，32*(10), 1736–1756.

Altfeld, S., Langenkamp, H., Beckmann, J., & Kellmann, M. (2017). Measuring the effectiveness of psychologically oriented basketball drills in team practice to improve self–regulation. *International Journal of Sports Science and Coaching, 12*(6), 725–736.

Bavard, S., Rustichini, A., & Palminteri, S. (2021). Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Science Advances, 7*(14), eabe0340.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 1–38.

Burton, D. (1989). Winning isn' t everything: examining the impact of performance goals on collegiate swimmers' cognitions and performance. *The Sport Psychologist, 3*(2), 105–132.

Campbell, E. M., Zhong, W., Hogeveen, J., & Grafman, J. (2025). Dorsal–ventral reinforcement learning network connectivity and incentive–driven changes in random exploration. *The Journal of Neuroscience, 45*(1), *e0422242025.*

Cunningham, J. B., & McCrum–Gardner, E. (2007). Power, effect and sample size using GPower: Practical issues for researchers and members of research ethics committees. *Evidence–Based Midwifery, 5,* 132.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model–based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215.

Ericsson, K. A., Krampe, R. T., & Tesch–Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363.

Feher da Silva, C., & Hare, T. A. (2018). A note on the analysis of two–stage task results: How changes in task structure affect what model–free and model–based strategies predict about the effects of reward and transition on the stay probability. *PLoS ONE, 13*(4), e0195328.

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model–based inference in the two–stage task. *Nature Human Behaviour, 4*(10), 1053–1066.

Feher da Silva, C., Lombardi, G., Edelson, M., & Hare, T. A. (2023). Rethinking model–based and model–free influences on mental effort and striatal prediction errors. *Nature Human Behaviour, 7,* 323–334.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal–directed choices by pruning decision trees. *PLoS Computational Biology, 8*(3), e1002410.

Jeong, Y. H., Healy, L. C., & McEwan, D. (2021). The application of goal setting theory to goal setting interventions in sport: A systematic review. *International Review of Sport and Exercise Psychology, 16*(1), 474–499.

Kevin, J. M., Carlos, D. B., & Matthew, M. B. (2016). *Identifying model–based and model–free patterns in behavior on multi–step Tasks.* bioRxiv.

Kingston, K. M., & Hardy, L. (1997). Effects of different types of goals on processes that support performance. *The Sport Psychologist, 11*(3), 277–293.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22*(6), 1027–1038.

Locke, E. (2000). Motivation, cognition, and action: An analysis of studies of task goals and knowledge. *Applied Psychology, 49*(3), 408–429.

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance, 3*(2), 157–189.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35–year odyssey. *American Psychologist, 57*(9), 705–717.

Locke, E. A., & Latham, G. P. (2013). *New developments in goal setting and task performance*. Routledge New York.

Louie, K., & De Martino, B. (2014). *The Neurobiology of Context–Dependent Valuation and Choice*. Elsevier.

Luna, R., Vadillo, M. A., & Luque, D. (2023). Model-free decision making resists improved instructions and is enhanced by stimulus-response associations. *Cortex, 168,* 102-113.

Luque, D., Molinero, S., Watson, P., López, F. J., & Le Pelley, M. E. (2020). Measuring habit formation through goal-directed response switching. *Journal of Experimental Psychology: General, 149*(8), 1449-1459.

Molinaro, G., & Collins, A. G. E. (2023). Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLoS Biology, 21*(7), e3002201.

Mullen, R., & Hardy, L. (2010). Conscious Processing and the Process Goal Paradox. *Journal of Sport and Exercise Psychology, 32*(3), 275-297.

Neiman, T., & Loewenstein, Y. (2011). Reinforcement learning in professional basketball players. *Nature Communications, 2*(1), 1283.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology, 68*(1), 73-100.

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications, 6*(1), 8096.

Rangel, A., & Clithero, J. A. (2012). Value normalization in decision making: Theory and evidence. *Current Opinion in Neurobiology, 22*(6), 970-981.

Rösch, D., Schultz, F., & Höner, O. (2021). Decision-making skills in youth basketball players: Diagnostic and external validation of a video-based assessment. *International Journal of Environmental Research and Public Health, 18*(5), 2331.

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems*.Cambridge University Press.

Sang, K., Todd, P. M., Goldstone, R. L., & Hills, T. T. (2020). Simple threshold rules solve explore/exploit trade-offs in a resource accumulation search task. *Cognitive Science, 44*(2), e12817.

Shteingart, H., Neiman, T., & Loewenstein, Y. (2013). The role of first impression in operant learning. *Journal of Experimental Psychology: General, 142*(2), 476-488.

Shteingart, H., & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology, 25*, 93-98.

Song, T., Ye, M., Teng, G., Zhang, W., & Chen, A. (2025). Expertise advantage of automatic prediction in visual motion representation is domain-general: A meta-analysis. *Psychology of Sport and Exercise, 76,* 102776.

Steffen, A., Rockstroh, B., Wienbruch, C., & Miller, G. A. (2011). Distinct cognitive mechanisms in a gambling task share neural mechanisms. *Psychophysiology, 48*(8), 1037-1046.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An Introduction.* The MIT Press.

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of Learning and Motivation* (pp. 37-76). Academic Press.

Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application, 8*(1), 1-23.

van Maarseveen, M. J. J., Savelsbergh, G. J. P., & Oudejans, R. R. D. (2018). In situ examination of decision-making skills and gaze behaviour of basketball players. *Human Movement Science, 57,* 205-216.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413-1432.

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018). Taxing the brain to uncover lying? Meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying. *Journal of Applied Research in Memory and Cognition, 7*(3), 462-469.

Wojciechowski, J., Olson, J. M., Subramanian, G., Kosowska, Z., & Pietras, K. (2025). The impact of reducing cognitive load in RT and P300 concealed information tests with importance related fillers. *International Journal of Psychophysiology, 209,* 112507.

# Process or Outcome? Context-Specific Decision Modeling of Basketball Players Based on Intrinsic Rewards

*Yang Liu [1], Zhang Xiyuan [2], Zhou Chenglin[1], Sang Biao[3]*

( [1] School of Psychology, Shanghai University of Sport, Shanghai, 200438 ) ( [2] China Basketball College, Beijing Sport University, Beijing, 100084 )

( [3] Laboratory of Educational Big Data and Decision-Making, Ministry of Education, Shanghai Institute of Educational Sciences, Shanghai, 200032 )

**Abstract** The decision-making process in basketball is inherently dynamic, involving the continuous integration of context and experience. While this process aligns with the core tenets of reinforcement learning (RL), classical RL models have been limited in their ability to capture the dynamic characteristics of multi-alternative, continuous decision-making. To address these limitations, the present study employed a modified version of the intrinsically enhanced model to investigate the effects of context and sports experience on the decision-making of basketball players. Moreover, the modified model was compared with several classic models, including one-trial back logistic regression, model-free model, and model-based model, to evaluate its performance in explaining and predicting decision behavior in the two-stage decision task.

This study aimed to examine how internal reward signals derived from process goals and external reward outcomes are integrated during decision-making in basketball. Specifically, we aimed to assess whether the incorporation of internal reward signals, which represent process-related achievements (e.g., tactical execution), could improve the predictive power of a modified intrinsically enhanced model. Furthermore, we explored how contextual factors modulate the decision strategies and reaction times, and whether these effects differ between experienced basketball players and novices.

A 2 (stimulus type: abstract symbols vs. basketball tactical diagrams) × 2 (group: novices vs. basketball players) mixed experimental design was employed. A total of 56 participants were recruited, including 29 basketball players with competitive experience and 27 novices with no basketball experience. Participants performed a two-stage decision task developed in MATLAB R2021b using Psychtoolbox (v3.0.19). At the outset of each trial, a fixation point was displayed for one second. In Stage 1 (S1), participants were presented with two options, representing either tactical initiation phases (in the basketball condition) or abstract alternatives (in the abstract condition). Their choice probabilistically determined the subsequent Stage 2 (S2) state, with common transitions occurring with a 70% probability and rare transitions with a 30% probability. During S2, participants made a binary decision and received immediate feedback (reward or no reward) with each option's reward probability fixed at 50%, thereby controlling for variability in external outcomes.

Behavioral responses and reaction times were recorded for both stages. Hierarchical Bayesian models were employed to analyze the data, with parameter estimation conducted via Bayesian methods using Stan and PyStan (v3.10) within a Python 3.12 environment. Model performance was evaluated using the widely applicable information criterion (WAIC), specifically the expected log pointwise predictive density (ELPD_WAIC), which simultaneously accounts for model fit and complexity. Key parameters of interest included the learning rate (α), which captures the effect of immediate reward feedback; the inverse temperature (β), which reflects the balance between exploitation and exploration; and the internal reward weight (θ), which indicates the degree to which process goals influence decision-making.

Model comparison revealed that the modified internally enhanced model outperformed all alternative models across all experimental conditions, as evidenced by consistently higher ELPD_WAIC values. This finding supports the enhanced model's ability to effectively integrate both internal reward signals and external outcomes in explaining decision behavior.

Analysis of the model parameters showed that, the α parameter, which reflects the immediate influence of reward feedback, remained stable across conditions. This finding implies that the underlying reward processing mechanism is robust and relatively unaffected by differences in experience or stimulus type. Although both basketball players and novices predominantly made outcome-driven decisions, basketball players exhibited a significantly higher sensitivity to process goals. Specifically, the internal reward weight (θ) for basketball players was approximately 0.33, indicating a relatively greater, but still sub-dominant, influence of process-related internal rewards compared to external outcomes. Furthermore, the β parameter was significantly lower in the basketball tactical diagram condition compared to the abstract condition, suggesting that participants were more inclined to employ an exploration strategy when exposed to contextually rich stimuli. Interestingly, the increased complexity inherent in the basketball tactical diagrams also led to prolonged reaction times. This finding indicates that additional cognitive load imposed by complex information requires individuals to invest more cognitive resources to process and integrate information, which tends to promote exploration strategies and prolong reaction times.

In summary, our study demonstrates that the modified intrinsically enhanced model provides a superior framework for capturing the dynamic decision-making processes of basketball players. While both experienced basketball players and novices primarily exhibit outcome-driven decision-making, basketball players display a higher sensitivity to process goals, reflecting the influence of extensive training and experience. Moreover, the increased information complexity associated with basketball tactical stimuli significantly prolongs reaction times and facilitates exploration strategies due to heightened cognitive load. These findings underscore the necessity of integrating both internal and external reward mechanisms to comprehensively model decision behavior in complex, real-world settings.

**Key words** reinforcement learning, intrinsically enhanced model, basketball, two-stage task, decision making