NBER 工作论文系列

典型入户辅导项目对儿童技能的影响

James J. Heckman

Bei Liu

Mai Lu

Jin Zhou

工作论文 27356

http://www.nber.org/papers/w27356

美国国家经济研究局（NBER）

1050 Massachusetts Avenue

Cambridge, MA 02138

2020 年 6 月，2022 年 2 月修订

人类发展经济学研究中心（CEHD）感谢新经济思维研究所和美国国立卫生研究院尤尼斯·肯尼迪·施赖弗国家儿童健康与人类发展研究所提供的资助，资助编号 R37HD065072。该项目已在 AEA 注册，注册号 AEARCTR-0007119。本文所表述观点仅代表作者个人，不一定代表资助方的观点或美国国立卫生研究院的官方观点。中国发展研究基金会（CDRF）感谢瑞银慈善基金会和敦和基金会的支持。本文作者衷心感谢 Susan Chang、Sally Grantham-McGregor、Sylvi Kuperman、Carey Cheng、Rebecca Myerson、Chunni Zhang 和 Yike Wang 在项目设计、项目实施和数据清洗等方面所做的努力。Erlfang Tsai 和 Fuyao Wang 在研究中提供了非常宝贵的协助。中国发展研究基金会对 Mary Young、Fan Bu、Peng Liu、Lijia Shi、Bojiao Liang 和 Yi Qie 表示感谢，他们为实地工作提供了重要和宝贵支持。参与者及其家人为该研究项目提供了大力支持，我们对此表示感谢。了解本文及其补充材料可访问 http://cehd.uchicago.edu/china-reach_home-visiting_appendix。本文所表述观点仅代表作者个人，不一定代表美国国家经济研究局的观点。

NBER 工作论文供讨论和评论之用，它们尚未经过同行评审，也未接受 NBER 董事会对其官方出版物进行的审查。

典型入户辅导项目对儿童基本技能的影响

James J. Heckman、Bei Liu、Mai Lu 和 Jin Zhou

NBER 工作论文，第 27356 号

2020 年 6 月，2022 年 2 月修订

JEL No. J13,Z18

摘要

本文采用随机分配方法，估算了一个广受效仿的儿童早期入户辅导项目对儿童基本技能的影响。我们展示了大规模复制该项目的可行性。我们估算了儿童个体的潜在技能向量，并对干预组和对照组做了比较。该项目大幅提升了儿童的语言和认知技能、精细运动技能和社交情感技能。我们不仅以未加权项目评分展示了干预效应，还查明了该项目是否会影响为一系列测验项目生成正确答案的潜在技能，及其如何影响从技能到项目评分的映射。潜在技能的增强可以解释大多数语言和认知方面的常规干预效应。该项目的主要运作方式是提高技能，而不是提高技能的使用效率。该项目几乎没有改变从潜在技能到项目测验评分的映射。

James J. Heckman

芝加哥大学人类发展经济学研究中心

1126 East 59th Street

Chicago, IL 60637

Mai Lu

中国发展研究基金会

中国北京东城区安定门外大街 136 号

皇城国际中心 A 座 15 层

和 IZA

以及 NBER

jjh@uchicago.edu

100011

lumai@cdrf.org.cn

Bei Liu

中国发展研究基金会

中国北京东城区安定门外大街 136 号

皇城国际中心 A 座 15 层

100011

liubei@cdrf.org.cn

Jin Zhou

芝加哥大学人类发展经济学研究中心

1126 East 59th Street

Chicago, IL 60637

jinzhou@uchicago.edu

数据附录载于 http://www.nber.org/data-appendix/w27356

1 简介

越来越多的研究表明，儿童早期入户辅导项目对于培养弱势儿童的技能非常有效。小规模入户辅导项目已被证明有效（参见 Howard 和 Brooks-Gunn，2009 年；HomVEE，2020 年；Grantham-McGregor 和 Smith，2016 年）。与许多其他儿童早期项目相比，其成本相对较低，而且它们在育婴辅导员培训和基础设施支持方面的要求也最低。育婴辅导员的教育水平与被探访者相当。牙买加入户辅导项目“Reach Up and Learn”启动于 30 多年前，其运作非常成功，引得世界各地纷纷效仿（Grantham-McGregor 和 Smith，2016 年）。

本文研究了中国西部贫困地区大规模开展“慧育中国”（以牙买加项目为蓝本）的情况（1500 多名参与者，而牙买加原项目的参与者为 100 多人）。该项目与牙买加原项目一样，通过随机对照试验进行评估。我们的证据表明，该项目进行大规模实施能够获得成功。

“慧育中国”项目比牙买加原项目的数据更丰富，部分原因在于这两个项目均由同一组学者设计，在此过程中他们还将牙买加经验融入了中国项目。我们的研究表明，该项目对语言和认知技能、精细运动技能和社交情感技能有着极大影响，但这种影响在基线分布中并不均匀。对于母亲不在身边的儿童来说，它在技能方面的积极影响最强烈。

为获得这些结果，我们背离了传统做法，调整了多个技能评估项目的任务难度级别。这样，我们避免了研究文献普遍采用的一个不合理方法——即对不同难度的任务进行非加权表现计分。通过调整方法，我们得出了更合理的预计干预效应。我们将预计干预效应分解为潜在技能的提高与技能运用能

力的提高。干预效应主要来自技能的提高。

本文结构安排如下：第 2 部分对该项目做出介绍，指出它是牙买加原项目的扩展版和增强版；第 3 部分介绍了一系列常规实验干预效应，并记录了项目影响的异质性。此外，我们估算了个人层面潜在技能的非线性因子模型，并确定了干预对可生成项目评分的技能的影响；第 4 部分研究了预计干预效应的来源。我们研究了该项目能在多大程度上影响将技能映射到任务表现的函数输入，以及它在多大程度上改变了固定潜在技能存量的生产力；第 5 部分比较了中国项目与牙买加原项目的结果（随访至 30 岁），“慧育中国”有望复制牙买加在教育和劳动力市场成果方面的长期成就；第 6 部分将就我们的调查结果做一总结。

2 “慧育中国”

正在实施的中国农村教育与儿童健康项目（“慧育中国”）启动于 2015 年，这迎合了中国国务院对循证式“从试点到政策”分析方法的日益重视与呼吁。此项大规模随机对照试验借鉴了牙买加的成功试点经验，旨在评估低成本入户辅导服务模式对弱势家庭的影响（参见 Grantham-McGregor 和 Smith，2016 年；Gertler、Heckman、Pinto、Zanolini、Vermeersch、Walker、Chang 和 Grantham-McGregor，2014 年）。该项目通过进一步扩大儿童与看护者及社区的接触面，来提高儿童的健康水平和认知能力。

甘肃华池县作为中国最贫困的地区之一，被选为该项目的实施地。华池县辖 15 个乡镇 111 个行政村，85%为山区，人口 13.2 万，其中农村人口 11.46 万。[1]从图 1 可看出，我们的研究项目启动于 2015 年 1 月，入户辅导开始于

[1]在中国，户口有农业户口和非农业户口之分，这种户籍登记制度定义并限制了国内人口的流动性。

2015 年 9 月。有关项目实施的详细信息请参阅附录 A。

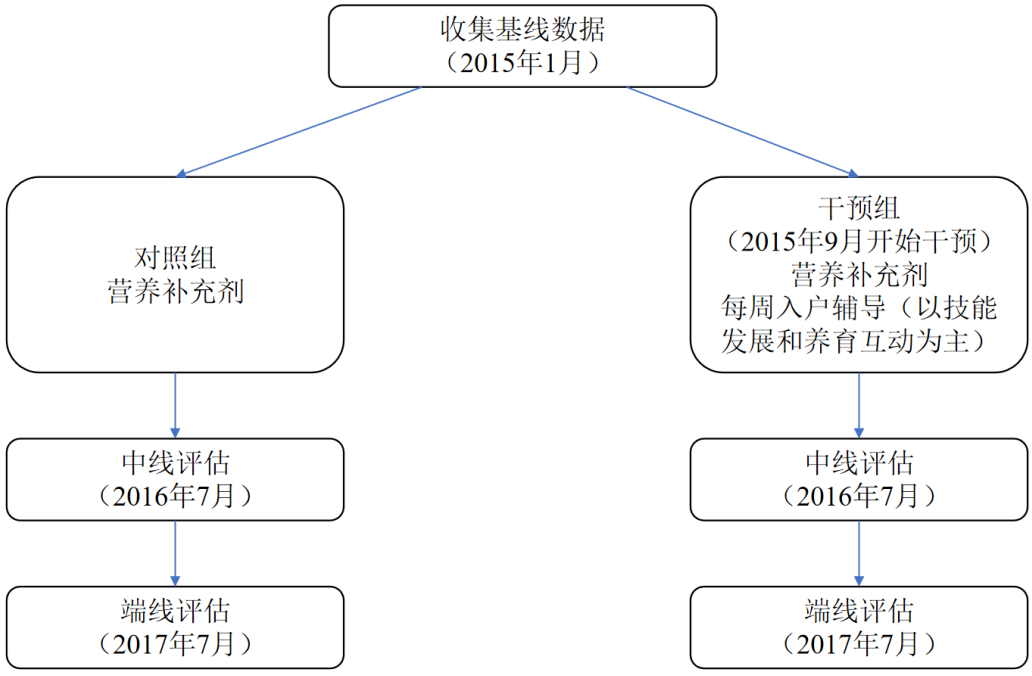


图 1：“慧育中国”（华池县）项目时间表

2.1 实施的干预措施

该项目选择了教育水平与受访母亲相当的人员进行培训，由他们入户辅导。这种做法在中国农村很容易复制，因为育婴辅导员的潜在供应量很大。该项目鼓励看护者选择适合儿童发展的方式，与儿童展开互动。Lizzeri 和 Siniscalchi（2008）开发了一种儿童发展模型，将亲子互动作为良好养育的重要决定因素。[2]

[2]Heckman 和 Zhou(2021)记录了所使用的入户辅导方案。

在这里，“慧育中国”项目由一名县总督导负责项目的整体协调，同时得到了 24 名乡镇督导员和 91 名育婴辅导员的协助。[3]协调人在全县范围内组织培训，对乡镇督导员进行监督。县总督导和乡镇督导员随机进行上门抽查，以便观察和了解育婴辅导员的工作情况。

督导员为育婴辅导员提供支持，并负责管理工作。他们会敦促育婴辅导员为每周入户辅导做好准备、回顾过去并计划未来的入户辅导活动，同时每周还要与育婴辅导员召开会议、进行反思，完善入户辅导项目和提升辅导体验。乡镇督导员每月与育婴辅导员一起入户一次，观察和记录看护者、儿童和育婴辅导员及其互动结果。

育婴辅导员每周都会走入家庭，按照牙买加项目方案提供一小时的育儿或看护指导和支持。[4]每次入户时，育婴辅导员都会记录有关家长参与的信息（如入户辅导时谁在照看儿童；如果儿童无法参加辅导，育婴辅导员是否教导父母开展相关任务；以及辅导后谁陪儿童一起玩耍，陪伴的频率是多少），以及儿童的表现情况（如上周布置的任务和本周的新任务）。Heckman 和 Zhou（2021 年）记录了“慧育中国”课程的内容、每周访问的内容以及每周使用的评估工具。该课程包括 200 多项与语言和认知技能发展相关的任务，其中有大约 70 项针对精细运动技能发展的任务和 20 项针对大运动技能发展的任务。

[3]乡镇是县的下属行政单位，平均每个育婴辅导员负责 8 户家庭。

[4]这些方案是以牙买加项目为基础，并借鉴中国文化做了调整(如将歌曲改为流行中文歌曲，添加中国人熟悉的背景图案)。针对 18 个月以下儿童的方案侧重于运动和语言技能训练。对于 18 个月以上的儿童，该方案增加了更多认知内容(如归类、配对和拼图)。

2.1.1 随机对照试验的设计

随机化是基于村一级集群的配对设计。Bai（2019 年）认为，这种设计最有利于确保平均干预效应估计值的均方误差最小化。实施过程分三步。我们首先核查了华池县所有符合条件的村庄。接下来，我们通过家庭调查并利用村一级行政数据，使用马哈拉诺比斯度量（居民和村庄特征）来评估村庄的相似性。[5]为确保每对儿中的马哈拉诺比斯度量最小化，我们按照度量评分对村庄进行排序，并使用非参数信念传播（nbp）匹配方法对最接近的村庄进行配对。[6]

对村庄进行配对后，我们从一对村庄中随机选择一个村庄作为干预组，另一个村庄作为对照组。[7]附录图 A.2 显示了华池县配对村庄的位置。这种设计紧密贴合了配对村庄的特点。[8]村一级干预效应包括了村内溢出效应。这些村庄作为干预组或对照组仅使用一次。

3 估算干预效应

“慧育中国”干预项目旨在促进多种技能（如运动、语言、认知和社交情感技能）的发展。表 1 显示了我们的技能衡量标准。Denver II 测验提供了

[5]用于匹配村庄对儿的村一级干预前协变量包括：(1)HOME IT 量表（家庭环境量表）中的“与儿童亲密程度”评分；(2)HOME IT 量表中的语言技能评分；(3)HOME IT 量表中的学习材料评分；(4)本村营养补充计划的实施率；(5)本村全县营养计划执行率；(6)儿童样本中留守儿童的占比；(7)本村人均纯收入；(8)本村平均受教育年限；(9)打算参与育儿干预项目的看护者占比；(10)打算携子女一同进城的家庭比例。

[6]Lu、Greevy、Xu 和 Beck(2011 年)。

[7] 配对村庄共有 55 对，也就是说干预组和对照组各有 55 个村庄。

[8] 附录 B 记录了基线比较。

详细的儿童发展评估任务措施。[9,10,11]

表 1：“慧育中国”入户辅导项目技能内容

技能类别	定义
精细运动	手指动作技巧，如抓、放、捏、画、写。
大运动	广泛的身体肌肉运动，如走、跑、扔、踢。
认知	学习技能，包括逻辑思维、解决问题、记忆力和注意力。
语言	发声、手势、说话连贯。
社交情感	以适当的方式表达和控制情绪、进行交流。

本部分报告了入户辅导干预对各类别内未加权项目评分的平均干预效应的常规估计值。项目评分是一项任务知识的二元指标。我们使用稳健的统计方法来调整缺失数据，并把村庄内部的干扰相互关联在一起（Cameron、Gelbach 和 Miller，2008 年）。

使用答对项目占比作为结果是一种标准做法，其假设是每个任务的测验难度级别相同。实际上，在我们的 Denver II 测验中任务难度级别存在很大差异。我们使用了一个非线性测量模型来解决这个问题，该模型考虑了项目难度（van der Linden，2016 年），并且恢复了产生项目响应的个人潜在技能。

[9]Denver II 测验专为临床医生、教师或儿童早期专业人士设计，旨在监测婴儿和学龄前儿童的发育情况。该测验主要是基于审查者的实际观察，而不是家长报告。该测验包含 125 项任务，期中包括四类技能测查：个人与社会(满足个人需求，与人相处)、精细运动适应性(手眼协调、操控小物件、解决问题)、语言(听力、理解和语言表达)和大运动(坐、走、跳和大肌肉整体运动)。

[10]附录 C 提供了中英文版的 Denver II 量表。

[11]Bayley III 能够将综合评分转换为基于年龄的量表评分，这在临床实践中更有用。不过，这一目标也可通过使用逐项 Denver II 测验来实现。Bayley III 测验针对的是 1 至 42 个月大的婴儿和儿童，其中包括审查者的观察(认知、运动和语言技能)，以及家长的问卷调查(社交情感和适应行为技能)。Ryu 和 Sim(2019 年)报告称，Denver 测验在检测语言发展延迟方面比 Bayley 测验更准确。

我们确定了实验引起的潜在技能改善状况，以及运用技能回答个别测验问题的改进情况。

3.1 县一级平均干预效应

我们现在对自己报告的干预效应做出定义。为便于说明，我们采用了一些符号。村庄使用{1,…,V}表示。村庄按照匹配规则 m(v)进行配对：$v\rightarrow\ v'$，其中根据干预前协变量均值向量$\bar{Z}\ (v)$，v'与 v 最接近。邻近度是通过马哈拉诺比斯度量来校准：

$$v'= \text{argmin}\ \left(\bar{Z}\ (v)\ -\ \bar{Z}\ (v')\right)'\sum\left(\bar{Z}\ (v)\ -\ \bar{Z}\ (v')\right)$$

$$\{1,...,V\}\backslash\{v'\}$$

其中 Σ 是按全部村庄计算的协方差矩阵 Z。我们采用掷币的方式来决定村庄对儿($v,\ v'$)中的哪个村庄接受干预，所有村庄仅选用一次。

如果v被选择进行干预，则记作 Dv =1。所有人 i 都被分配到某个村庄。Dv(i)是 i 在v中指定的干预状态，Dv(i) $\in$ {0, 1}。每个村庄都有 Iv个符合资格的居民。

我们首先报告了对标准化评分的平均干预效应，此类评分是根据以下实证模型估算而来：

$$\Upsilon_{iv}^{m} = \beta_0 + D_{v(i)}\beta_1^m + Z_i'\beta_2^m + \sum_{p=1}^{P} 1\{i \in p\}+\varepsilon_{iv}^{m} \qquad (1)$$

其中Υ_{iv}^{m}是村庄v中儿童i的结果m的标准化评分，D$v(i)$ 是一个虚拟变量，表示儿童i所在村庄v的干预状态，Z_i是干预前协变量。1{$i \in p$}是儿童i是否居住在村庄对儿p.$\Upsilon_{iv}^{m} = D_{v(i)}\Upsilon_{iv}^{m}(1) + (1 - D_{v(i)})\Upsilon_{iv}^{m}(0)$中的指标，其中

$\Upsilon_{iv}^{m}(d)$表示固定干预状态d的结果向量。干预分配设计意味着

$$\left(\Upsilon_{iv}^{m}(0),\Upsilon_{iv}^{m}(1)\right)\coprod D_{v(i)} \mid Z_i. \tag{2}$$

干预是在村一级进行。儿童i 的异质冲击项ε_{iv}^{m}可任意与同一村庄v中的其他任何儿童$i' \neq i$的ε_{iv}^{m}关联在一起。假设异质冲击在各个村庄间是独立的；即在$i \in v$和$\forall k \in v'$, $v \neq v'$时为$\varepsilon_{iv}^{m}\coprod\varepsilon_{kv'}^{m}$。附录 D 显示的残差图验证了各村庄残差独立性的假设。$N \times N$ 协方差矩阵 $E(\epsilon\epsilon') = \Omega$ 与村庄数 V 呈块对角线关系：$\Omega vv' = 0$；所有 $v \neq v'$.[12]

随着每个集群中观测值数变大，以及集群数增多，同时假设集群与集群中观测值的比率为一个常数时，则参数(1)的 OLS 估计量一致。如果β_1^m在人群中为恒定，则这是正确的。

通过X_{iv}定义(1)中右侧变量的完整阵列。当使用 OLS 残差$\hat{\epsilon}_v$: $E(\hat{\epsilon}_v\hat{\epsilon}_v')$估算$\hat{\Omega}v$时，标准集群稳健方差估计量(CRVE), $(X'X)^{-1}(\sum_{v=1}^{V} X_v' \hat{\Omega} vXv)(X'X)^{-1}$，会出现偏差。[13]这种偏差取决于 Ωv的形式。Cameron、Gelbach 和 Miller（2008 年）讨论了这个问题，并指出野生自助法在进行集群稳健性推断方面表现良好。有关我们使用的野生自助法详细信息请参阅附录 E。[14]

在我们的样本中，接受干预的村庄中 98%以上符合条件的儿童都受到了入户辅导。尽管如此，对照组和干预组仍有约 15%的儿童错过了年度儿童发育评估。为对总体平均干预效应做出一致估算，我们使用了逆概率加权

---

[12] X_v表示第 v 个集群中的 X，且 $E(\epsilon v) = 0$, $E(\epsilon v\epsilon_v')$ = Ωv。X 包括匹配对儿的干预状态、干预前协变量和指标。

[13] $\hat{\epsilon}_v$是 OLS 残差。

[14]由于我们有 55 个集群，因此最近对原始聚类自助法的担忧不会出现在这里。参见 Canay、Santos 和 Shaikh(2019 年)。

（Tsiatis，2006 年）。[15,16]

表 2 使用标准化结果测量工具得到了对每个技能类别的干预效应。[17,18]我们采用了不同的统计模型，第(1)、第(2)和第(3)列使用了所有可用数据样本，第(4)和第(5)列仅使用了 2015 年 9 月项目启动时 2 岁以下儿童的样本。年龄较小的受干预儿童至少接受了一年干预。[19]

表 2 的第一行显示，平均而言，干预组儿童更有可能获得较高的语言和认知技能。[20]在第一行中，我们看到在中线处（干预约九个月后），干预组儿童的语言和认知技能比对照组儿童高出约 0.7 个标准差。干预结束时，对语言和认知技能的干预效应超过了 1 个标准差。干预显著提高了受干预儿童的语言和认知技能。当干预组的儿童能够较早、较长时间接触育婴辅导员时，年龄调整干预效应的幅度增大（参见第(4)和第(5)栏）。这与动态互补性是一致的。

干预显著提高了中线处的社交情感技能和干预结束时的精细运动技能，但没有显著改善大运动技能。这一发现与课程设计保持了一致，其主要侧重

[15] Maasoumi 和 Wang(2019 年)使用 IPW 方法进行了稳健性推断，来剔除低概率观测值。在我们研究中，只有三个观测值的倾向评分(非缺失)低于 0.1，因此我们无需裁剪数据就能避免不一致的问题。

[16]附录 F 记录了数据损耗问题以及我们如何计算丢失数据的概率。为避免冗余，我们在所有估算中包含了逆概率。

[17]只有 140 名儿童参加了 Denver 基线测验。经过对拥有基线信息的儿童进行相同的模型估算，我们没有发现对照组和干预组在 Denver 测验评分上存在显著差异。有关此平衡测验的详细信息请参阅附录 B。

[18]Denver 测验在中国还没有总体层面的参考值。我们将对照组作为参照组：我们按月龄估算了 Denver 测验情况，然后使用均值和方差来制定干预组和对照组每个月龄组的测验评分标准。

[19]限制样本有两个原因。(1)正如本文所言，我们希望干预组的儿童能大量接受干预。许多年龄较大儿童参与的时间较短。(2)对照组中年龄较大儿童多于干预组，这是因为现场研究团队在 2015 年 9 月后没有更新干预组的名单。

[20]我们将这些类别组合在一起，来获得一些与其他类别相当的项目评分。

于发展语言和认知。[21,22]

表 3-4 显示了按性别划分的干预效应。一个有趣的发现是干预对男孩语言和认知技能的改善效果远超过女孩，这与文献中反复出现的结果一致（Elango、García、Heckman 和 Hojman，2016 年）。在中线处，女孩的干预效应量为 0.4 个标准差，男孩的干预效应量为 0.9 个标准差。干预结束时，女孩的效应量约为 0.9 个标准差，男孩的效应量约为 1.1 个标准差。这里的原因之一是在儿童早期阶段，女孩平均比同龄男孩发育得更成熟。干预组中的女孩在社交情感技能方面也有更好表现。[23]

[21]Heckman 和 Zhou(2021 年)记录了干预课程。

[22]在使用原始评分而不是标准化评分时，结果具有可比性。有关这些报告请参阅附录 D。

[23]佩里学前教育项目(Heckman 和 Karapakula，2019 年)和初学者学前项目(García、Heckman 和 Ziff，2018 年)的评估中也出现了这一结果。

表 2：以 Denver 标准化评分为结果变量的干预效应

	(1)	(2)	(3)	(4)	(5)
	全部	全部	全部	参与时≤2岁的儿童	参与时≤2岁的儿童
			中线		
语言和认知	0.589***	0.631***	0.714***	0.674***	0.741***
	[0.234, 0.965]	[0.237, 1.036]	[0.319, 1.093]	[0.279, 1.067]	[0.350, 1.144]
精细运动	0.334	0.559	0.633*	0.629*	0.703*
	[-0.140, 0.787]	[-0.032, 1.174]	[0.003, 1.313]	[0.023, 1.324]	[0.057, 1.375]
社交情感	0.690**	0.865***	0.879***	0.624***	0.620***
	[0.260, 1.117]	[0.421, 1.312]	[0.467, 1.289]	[0.129, 1.118]	[0.204, 1.067]
大运动	-0.051	-0.004	-0.015	0.054	0.010
	[-0.598, 0.478]	[-0.564, 0.577]	[-0.567, 0.554]	[-0.514, 0.640]	[-0.559, 0.584]
			端线		
语言和认知	0.979***	0.914***	1.036***	1.016***	1.113***
	[0.585, 1.402]	[0.495, 1.347]	[0.644, 1.458]	[0.637, 1.408]	[0.723, 1.510]
精细运动	0.585**	0.574**	0.676***	0.561**	0.645**
	[0.006, 0.956]	[0.067, 1.091]	[0.180, 1.170]	[0.030, 1.095]	[0.139, 1.158]
社交情感	-0.201	-0.276	-0.222	-0.167	-0.115
	[-0.596, 0.202]	[-0.688, 0.123]	[-0.636, 0.194]	[-0.553, 0.215]	[-0.491, 0.275]
大运动	0.067	0.125	0.173	0.155	0.219
	[-0.479, 0.632]	[-0.392, 0.645]	[-0.322, 0.668]	[-0.406, 0.732]	[-0.294, 0.775]
干预前协变量	No	No	Yes	No	Yes
IPW	No	Yes	Yes	Yes	Yes

表 3：以 Denver 标准化评分为结果变量的干预效应

（女性）

	(1) 全部	(2) 全部	(3) 全部	(4) 参与时≤2岁的儿童	(5) 参与时≤2岁的儿童
			中线		
语言和认知	0.410	0.417	0.445	0.511**	0.534**
	[-0.076, 0.869]	[-0.035, 0.884]	[-0.014, 0.910]	[0.040, 0.991]	[0.080, 0.990]
精细运动	0.400	0.399	0.335	0.512	0.544
	[-0.252, 1.049]	[-0.271, 1.065]	[-0.269, 1.211]	[-0.088, 1.142]	[-0.082, 1.189]
社交情感	1.020***	1.068***	1.114***	0.912**	0.938***
	[0.445, 1.614]	[0.520, 1.614]	[0.681, 1.550]	[0.272, 1.541]	[0.400, 1.431]
大运动	0.117	0.063	0.058	0.085	0.019
	[-0.487, 0.751]	[-0.565, 0.665]	[-0.532, 0.675]	[-0.514, 0.725]	[-0.605, 0.652]
			端线		
语言和认知	0.852**	0.895**	0.950**	0.865**	0.893**
	[0.077, 1.596]	[0.159, 1.612]	[0.213, 1.675]	[0.122, 1.590]	[0.177, 1.598]
精细运动	0.804**	0.815**	0.866**	0.836**	0.855**
	[0.111, 1.500]	[0.088, 1.553]	[0.189, 1.574]	[0.110, 1.554]	[0.117, 1.579]
社交情感	-0.264	-0.298	-0.309	-0.264	-0.291
	[-0.806, 0.254]	[-0.805, 0.267]	[-0.775, 0.160]	[-0.859, 0.342]	[-0.820, 0.206]
大运动	0.188	0.246	0.257	0.460	0.445
	[-0.737, 1.091]	[-0.668, 1.094]	[-0.582, 1.080]	[-0.410, 1.308]	[-0.417, 1.326]
干预前协变量	否	否	是	否	是
IPW	否	是	是	是	是

表 4：以 Denver 标准化评分为结果变量的干预效应

（男性）

	(1) 全部	(2) 全部	(3) 全部	(4) 参与时≤2 岁的儿童	(5) 参与时≤2 岁的儿童
			中线		
语言和认知	0.747*** [0.236, 1.257]	0.852*** [0.261, 1.462]	0.938*** [0.389, 1.499]	0.896*** [0.345, 1.460]	0.911*** [0.329, 1.501]
精细运动	0.395 [-0.108, 0.908]	0.674 [-0.083, 1.532]	0.716 [-0.099, 1.598]	0.730 [-0.028, 1.577]	0.771 [-0.070, 1.747]
社交情感	0.436 [-0.115, 0.989]	0.589* [0.028, 1.140]	0.549** [0.047, 1.054]	0.395 [-0.178, 0.946]	0.280 [-0.272, 0.842]
大运动	-0.066 [-0.798, 0.661]	0.079 [-0.728, 0.900]	-0.041 [-0.700, 0.639]	0.152 [-0.634, 0.963]	-0.021 [-0.682, 0.659]
			端线		
语言和认知	1.050*** [0.514, 1.560]	0.797** [0.205, 1.436]	0.950*** [0.448, 1.497]	1.000*** [0.468, 1.513]	1.111*** [0.625, 1.626]
精细运动	0.460 [-0.212,1.117]	0.388 [-0.314, 1.108]	0.462 [-0.206, 1.144]	0.346 [-0.374, 1.042]	0.388 [-0.355, 1.124]
社交情感	-0.139 [-0.643, 0.390]	-0.306 [-0.895, 0.305]	-0.256 [-0.829, 0.326]	-0.157 [-0.654,0.351]	-0.169 [-0.701, 0.400]
大运动	-0.059 [-0.528, 0.424]	-0.071 [-0.543, 0.407]	-0.048 [-0.510, 0.419]	-0.169 [-0.663, 0.332]	-0.138 [-0.629, 0.359]
干预前协变量	否	否	是	否	是
IPW	否	是	是	是	是

注： 1. 括号内 95%置信区间是在村一级使用原始聚类自助法进行的估算。

2.标准化评分的均值和方差是根据对照组儿童的汇总样本来估算。

3.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.调整项目难度后，对社交情绪技能的负面干预效应就会消失。

5.“全部”列包含了所有观测值，“参与时≤ 2 岁的儿童”列将样本限制为参与该项目时年龄在 2 岁以下的儿童。

附录 G 分析了育婴辅导员与看护者之间、育婴辅导员与儿童之间的互动对儿童基本技能的影响，并展示了可反映育婴辅导员教学能力的变量。[24]结果显示，唯一的强效果是看护者与育婴辅导员之间产生良好互动，可促进儿童的语言和认知技能。[25]

[24]互动效果每月都会记录。用于中线回归的测量值是中线处每月测量值的均值。用于端线回归的测量值是整个干预期间测量值的均值。

[25]附录 G 中，表 G.2 显示出这些测量值具有极大分散性，因此交互效应的弱估计并不是由于样本方差不足造成的。

3.2 考虑任务难度因素，对儿童潜在技能的干预效果的估计做调整

先前分析表明，干预可提高未加权内容项总量的结果。如此形成的聚合虽然传统，但除非不同内容项的难度相同，否则就会产生问题，从评估的设计结果来看这是不正确的。

为解决这个问题，我们利用数据的多内容项的特点，估算了个人层面潜在技能的非线性因子模型。[26]我们遵循心理测量学的标准方法，引入并估算了跨内容项的难度参数（van der Linden，2016 年）。我们还评估了个人层面潜在技能。我们使用这些估计值来确定干预对产生项目评分的技能的影响。我们还估算了干预在多大程度上能改变技能和项目评分之间的映射关系（即受干预儿童能否更好地利用现有技能）。

3.2.1 项目和技能

我们研究的结果是儿童在个人任务上的表现，这是通过他们在测验项目上的表现来衡量。每个特定技能 K 都有 N_{Jk} 个任务，这些任务都是针对某一技能（如运动、认知、阅读等）。假设这些任务上的表现是由潜在技能 θ 产生。我们使用 NJ 来表示所有技能的项目总数（即 $N_J = \sum_{k=1}^{k} N_{Jk}$）。我们假设所有村庄都采用了一种将技能映射到测验评分的通用技术，因此我们放弃了特定于 v 的表示法。设Υ_i^{jk}(d)为二元值结果变量，表示人员 i 对技能类型 k 中任务 j 的掌握情况。对于干预状态 $d \in \{0, 1\}$的人来说，其表现是由任务项 j 的潜在结果生成。设θ_i^d是具有干预状态 d 的人的潜在技能 K 维向量，X_i 是基线协变量的一个向量。把从潜在技能θ_i^d到任务 j 的结果的决定因子映射关系写

[26]在数据中，我们为每个人的各项技能提供了 70 多个项目，用于衡量 Denver 测验中的任务表现。

为

$$\tilde{\Upsilon}_i^{jk}(d) = X_i'\beta^{jk,d} + \delta^{jk} + (\theta_i^d)'\alpha^{jk,d} + \varepsilon_i^{jk}, j = 1, \ldots, N_{Jk}; k = 1, \ldots, K.$$

$$\Upsilon_i^{jk}(d) = \begin{cases} 1\ \tilde{\Upsilon}_i^{jk}(d) \geq 0 \\ 0\ \tilde{\Upsilon}_i^{jk}(d) < 0 \end{cases} \tag{3}$$

其中 $\alpha^{jk,d}$ 是因子载荷的 K 维向量；δ^{jk} 是任务项 j_k 的任务难度参数；系数 $\beta^{jk,d}$ 和 $\alpha^{jk,d}$ 可取决于干预、建模技能甚至所研究的项目（项目在人群中是共同的）。在估算中，我们设 $\beta^{jk,d} = \beta^{j'k',d} = \beta^{k,d}$, $\forall jk$ 和 $j'_{k'}$；即系数在技能内的各个项目之间是通用的。

该模型将干预解释为影响任务表现的可塑造技能。干预还可提高执行任务时任何特定技能的生产力，即干预转移 $\alpha^{jk,d}$。对于任一来源的干预 $D=\mathrm{d}$ 的结果 jk 来说，对象$(\theta_i^d)'\alpha^{jk,d}$是一组有效技能。

在适当标准化下，我们可以识别个人层面的潜在技能因子θ_i^d，而不仅仅是传统心理测量模型中潜在技能因子的分布情况（如参见 van der Linden，2016 年）。我们假设ε_i^{jk}是单位法向量，独立于其他右侧变量。该数据在项目上具有类似于面板的结构，它可使用带有潜在技能的概率模型进行拟合。我们估算了可观察协变量的参数、潜在因子以及潜在技能因子对结果的影响。从 Wang（2020 年）的分析可以看出，当观测数（样本参与者）$N_I \to \infty$和 $N_J \to \infty$时，模型参数（包括个人能力）的估计量具有一致性且渐近无偏，但$\frac{N_I}{N_J}$会收敛于一个常数。[27]这些条件适用于我们的样本：每个人有大量的测验项目和观察值。

想把 θ^d 从$\alpha^{jk,d}$中分离出来，就要对因子模型进行标准化处理。由于

[27]回想一下，在估计中项目的数量可根据实际测验设计而变化。

$\theta_i^{d'}\alpha^{jk,d} = (\theta_i^d)'AA^{-1}\alpha^{jk,d}$，因子和因子载荷本质上具有任意性，除非以某种方式设定占比。如果满足于衡量有效技能$\theta_i^{d'}\alpha^{jk,d}$的变化，我们就能避免这种标准化。我们可使用 Anderson 和 Rubin（1956 年）提出的标准化概念来分解这个术语，并识别向量θ_i^d和$\alpha^{jk,d}$。我们分别报告了θ_i^d和$\alpha^{jk,d}$ 的估计值，也将其作为一组有效技能$(\theta_i^d)'\alpha^{jk,d}$进行了报告。

按照 Rasch 模型文献（van der Linden，2016 年）中的传统，我们假设 δ^{jk} 是测量系统所固有的干预不变任务难度参数，并且与干预状态无关。这确保了干预和对照之间测量结果的可比性。

我们的模型中有四种不同潜在技能因子，分别对应于 Denver II 测验 $k \in \{1,\ldots,4\}$中的社交情感、语言和认知、精细运动和大运动技能。为解释这些因子，我们假设 N_J 任务$(K \leq N_J)$ 中的 K 任务表现仅取决于一个因子。这就是 Cunha、Heckman 和 Schennach（2010 年）所说的“专用因子案例”，仅适用于每次测量的前四个项目。因此，我们在概括他们的分析时，要求只将一部分任务用于各类技能测量。我们对因子载荷矩阵进行了标准化，以便前 K 行构成一个 $I_{K,K}$ 单位矩阵。对于前 $K=4$ 个测量项目，我们假设它们仅加载一项技能。[28]N_J 结果向量的剩余因子载荷矩阵不受限制。剔除 d 上标来减少符号混乱后，我们将潜在技能的载荷指标书写为$\alpha'_{N_J\times K}$:

[28]我们选取了洗手并擦干项目、模仿垂线项目、组词项目和跳远项目，来分别展示社交情感技能、精细运动技能、语言和认知技能以及大运动技能。中国非常重视营造卫生、安全的社会环境，因此洗手并擦干是一项重要的社交技能。

$$\alpha'_{N_J\times K}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ \alpha^{5,1} & \alpha^{5,2} & \alpha^{5,3} & \alpha^{5,4}\\ \vdots & \alpha^{6,2} & \cdots & \cdots\\ \alpha^{N_J,1} & \cdots & \cdots & \alpha^{N_J,4}\end{bmatrix} \tag{4}$$

我们检验并拒绝了“专用模型”，该模型假设在(4)中的j_k行内，除了一个 $\ell \in \{1,\ldots,4\}$外，当$j_k \geq 5$时$\alpha^{jk,\ell,d}=0$。表 5 展示了该测验。专用因子模型的假设在我们样本中并不成立。

表 5：除了一个 $\ell \in \{1,\ldots,4\}$外，当$j_k \geq 5$时$\alpha^{jk,\ell,d}=0$的假设检验

	对照组		干预组	
	$\mathcal{X}^2(68)$	p 值	$\mathcal{X}^2(68)$	p 值
社交情感	463.247	0.000	1434.742	0.000
精细运动	494.200	0.000	1418.862	0.000
语言和认知	1186.793	0.000	2108.501	0.000
大运动	1570.322	0.000	1969.099	0.000

在附录 H 中，我们使用各种似乎可信的标准化工具对相关估算进行了敏感性分析。我们发现，文中报告的$\alpha^{jk,d}$估算在各种标准化下是稳定的。[29]我们的结果在数量上具有稳健性。我们利用 Chen、Fernández-Val 和 Weidner（2021 年）提出的估算程序，估算了具有多个潜在技能因子的面板概率模型。[30]采用这种方法估算个体特定因子和总体因子载荷的渐近理论依据来自 Wang（2020 年）。

3.2.2 估计值

[29] 在附录 H 中，我们比较了不同标准化下技能载荷的分布。我们发现，当选择中等难度级别的项目时其结果具有稳健性。

[30]有关该方法的详细信息请参阅附录 I。

表 6 给出了 $\beta^{k,d}$ 的估计值。尽管干预组中的男性点估计的负值明显更大，但干预组和对照组之间没有在统计学上表现出显著差异。图 2 比较了我们模型中所预测组合语言和认知任务项与实际任务项的分布情况。[31]我们还将数据与其他类型的任务很好拟合在了一起。[32]

表 6：可观察协变量的系数估计值

	对照组	干预组
月龄	0.961	0.924
	[0.166, 1.987]	[0.161, 1.738]
月龄 2	-0.009	-0.009
	[-0.025, 0.002]	[-0.0193, 0.002]
男性	0.356	-0.144
	[-1.081, 2.363]	[-1.178, 1.148]
常数	-16.756	-15.571
	[-35.260, -2.727]	[-31.620, -2.457]
	$x^2(4) = 0.004$	$p = 0.999$

注：1.括号内为 95%置信区间。

2.置信区间由村一级的成对聚类自助法构设而成。

3.我们使用概率比检验来检查两组的系数是否相同。检验结果表明，我们不能拒绝这些系数相同的假设。

[31]由于 Denver 测验中少有认知测验项目，因此我们将语言和认知任务合并为一类。

[32]参见附录 J。

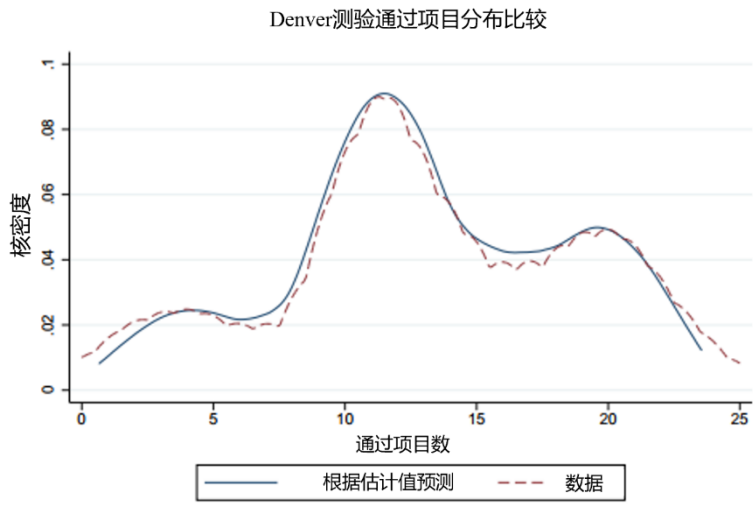


图 2：Denver 测验通过项目分布

图 3 显示了为每个任务项估算的难度级别参数 δ^{ik}。当项目难度增加时，估计值会变得更消极。这些估算大体上符合测验的设计结果，即增加后面项目的难度。估算的难度级别参数 $\delta^{ik}k$ 提供了有关测验设计是否合理的信息。比如，大运动技能的测验设计就不是特别好：难度级别数值在-1.8 左右持平，然后在第五项时迅速跳至-6。这意味着参加测验的儿童可以答对简单的问题，但无法回答所有较难问题。与大运动技能任务项相比，语言和认知任务项的设计显得更好，所有项目难度水平都在平稳上升。然而，社交情感任务项的估计值并不符合预期的评估设计结果。

表 7：对潜在技能因子均值的干预效应

	社交情感	精细运动	语言和认知	大运动
干预组	0.395***	0.726***	0.753***	-0.095
	[0.208, 0.583]	[0.551, 0.899]	[0.459,1.051]	[-0.280, 0.089]

注：1.括号内 95%置信区间是在村一级使用原始聚类自助法进行估算。

2.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

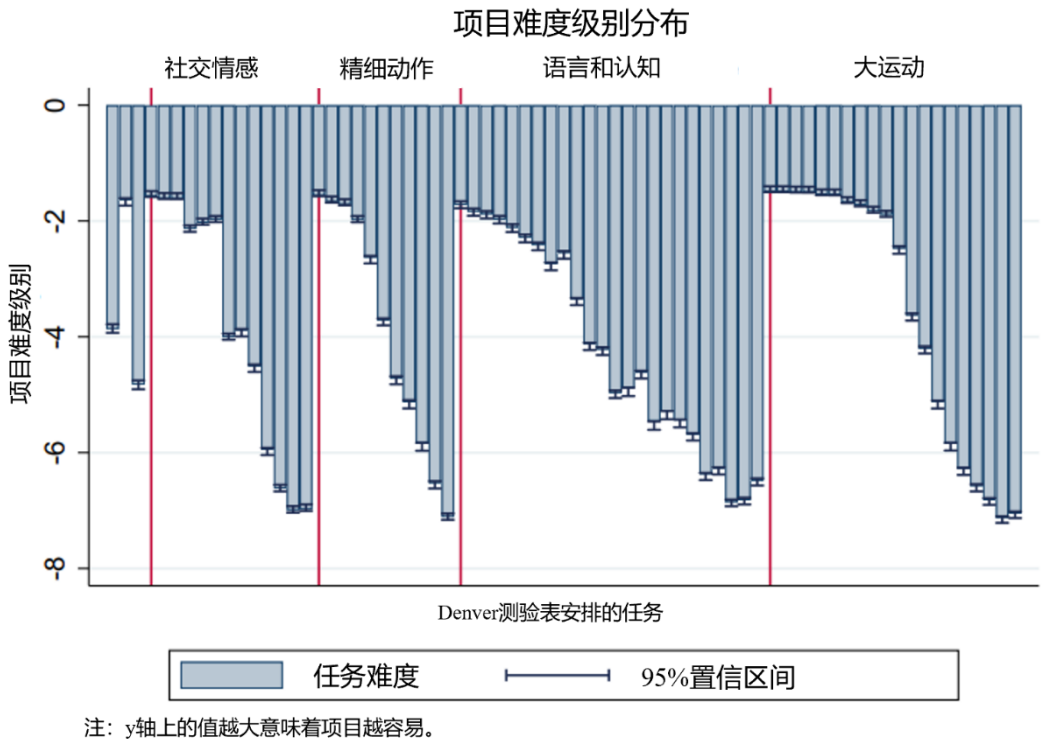


图 3：Denver 任务项难度级别分布

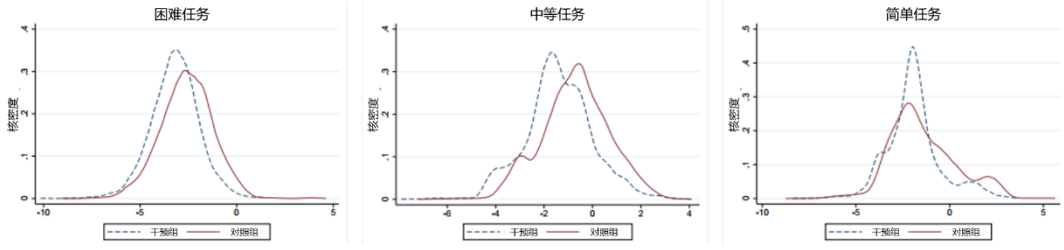
表 8：不同潜在技能因子之间的相关性

	社交情感	精细运动	语言和认知	大运动
社交情感	1			
精细运动	0.428***	1		
语言和认知	0.455***	0.207***	1	
大运动	0.085***	0.156***	-0.102***	1

注：1.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

我们方法的一个优点是可以估算个人层面的潜在技能因子。首先，表 7 列出了对四种潜在技能因子均值的干预效应。除大运动技能外，干预组所有

其他潜在技能因子的均值均显著高于对照组。在比较不同潜在技能的干预效应时，我们发现精细运动和语言技能获得了相同改善，但大运动技能没有受到影响。从表 8 可看出，语言和认知技能与大运动技能呈负相关，与社交情感和精细运动技能呈正相关。



(1)困难任务　　(2)中等任务　　(3)简单任务

图 4：$\left[(\theta_i^d)'\alpha^{jk,d}\right]^+$的分布

+ 共有 72 个任务按难度排序。简单任务是难度参数排名在 1~24 之间的任务，中等任务是难度参数排名在 25~48 之间的任务，困难任务是难度参数排名在 49~72 之间的任务。

图 4 给出了基于 Denver 任务难度级别的估算技能因子载荷与潜在技能因子乘积。[33]干预组在困难任务和中等任务中的载荷较大，而在简单任务中较小，这表明较容易的任务无助于检测干涉对儿童基本技能发展的干预效应。在其他技能方面，干预组和对照组的载荷模式相似。有关载荷总量的估算是精确的，对于大多数任务我们拒绝$\alpha^{jk,\ell,d=1}$ $\alpha^{jk,\ell,d=0}, \ell \in \{1,\ldots,4\}$的假设。[34]唯一的强相关性是社交情感技能和精细运动技能之间的相关性。

[33] 附录 I 介绍了其他类型任务的潜在技能载荷。我们总共有 72 个任务，便将难度参数最高的 24 个任务设为简单任务，最低的 24 个设为困难任务，中间的 24 个设为中等任务。所有等级都是基于对任务难度级别参数的估计。

[34]附录 H 中，表 H.3 – H.4 提供了逐项测验结果。社交情感项目载荷无法进行精确估算。

表 9：Denver 测验任务($\alpha^{jk,d}$)潜在技能的技能载荷

对照组			干预组			p 值
技能载荷	均值	标准差	技能载荷	均值	标准差	用于均等检验的 p 值
语言和认知	0.453	0.364	语言和认知	0.679	0.469	0.000
社交情感	0.259	0.263	社交情感	0.222	0.246	0.002
精细运动	0.448	0.251	精细运动	0.556	0.211	0.001
大运动	0.739	0.405	大运动	0.693	0.442	0.276

注：1.这些分别是各项目中的 $\alpha^{jk,0}$ 和 $\alpha^{jk,1}$ 均值和标准差。

2. p 值是对原假设干预组和对照组总体测量值相等的显著性检验。

从等式(3)可看出，在相同技能水平下因子载荷越大，儿童在测验中表现就越好。表 9 给出了不同任务下技能载荷的汇总统计数据（均值和标准差）。除大运动技能外，我们拒绝干预组和对照组的汇总统计数据平等。此外，该表还显示了每种技能在实施各种任务时的平均有效性。比如，潜在语言和认知技能在语言和认知任务中的载荷较大，而社交情感技能在语言和认知任务中的载荷相对较小。这让我们对所采用的标准化有了一些信心。

3.2.3 与无任务难度参数的模型相比较

为显示将任务难度参数引入模型的影响，我们根据等式(3)估算了模型的限制版本，将所有任务难度参数设置为零。首先，我们比较了全模型和受限模型之间的概率比，发现全模型的概率比更高。概率比检验统计量为 $\chi^2(71) = 8419.26$，基于两个模型拒绝拟合优度相同零假设的 p 值小于 0.001。

其次，我们在表 10 中比较了对潜在技能因子均值的干预效应($E(\theta^1) - E(\theta^0)$)。请注意，无任务难度参数模型的估计值与有难度参数模型的估计值有很大差异。无难度参数模型对社交情感技能有着显著的负面影响，而对大运动技能有着显著的正面影响，这与全模型和 OLS 模型干预效应评估不一致。

表 10：根据有无难度参数比较两种模型的干预效应 θ_i

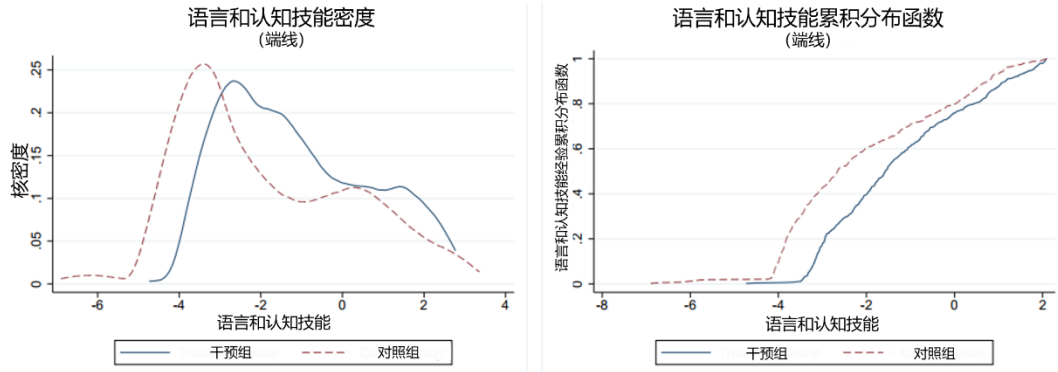
	社交情感	精细运动	语言和认知	大运动
全模型	0.395***	0.726***	0.753***	-0.095
（调整任务难度）	[0.208, 0.583]	[0.551, 0.899]	[0.459, 1.051]	[-0.280, 0.089]
受限模型	-3.14***	1.136***	1.158***	1.069***
（没有调整任务难度）	[-3.375, -2.904]	[1.205, 1.505]	[0.857, 1.453]	[0.896, 1.237]

注：1.括号内 95%置信区间是在村一级使用原始聚类自助法进行的估算。

2.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2.4 潜在技能分布

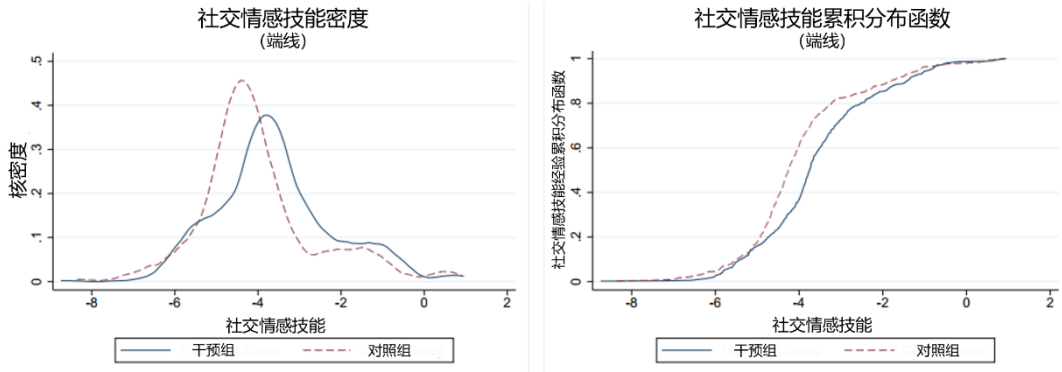
我们比较了对照组和干预组的语言和认知技能分布。图 5a 显示，干预组的语言和认知技能密度向右移动，并且上侧尾部比对照组更宽。图 5b 显示，几乎在累积分布的每个点上，干预组的语言和认知技能都高于对照组。与处于控制分布顶部的儿童相比，处于控制分布底部和中部的儿童收益更可观。



(1) (2)

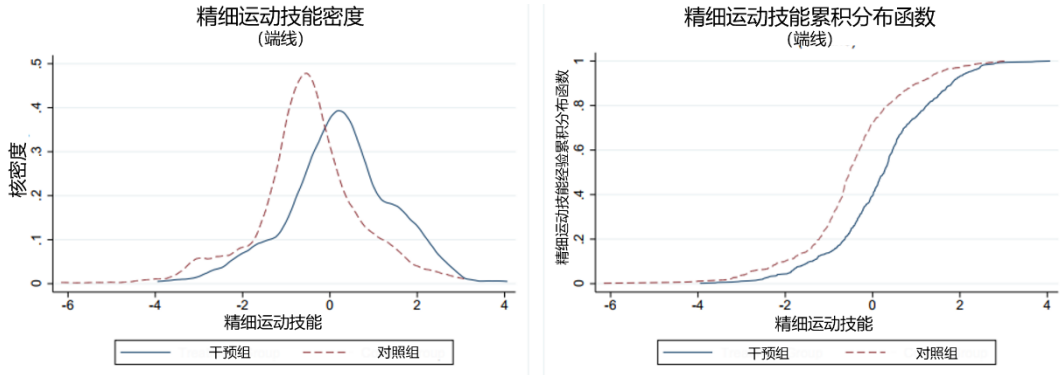
图 5：语言和认知技能分布

图 6a 和 7a 分别显示了社交情感和精细运动技能的密度。对于社交情感技能，收益主要集中在那些本应处于控制分布中心的儿童身上。对于精细运动技能，整个控制分布中的收益都很可观。



(1) (2)

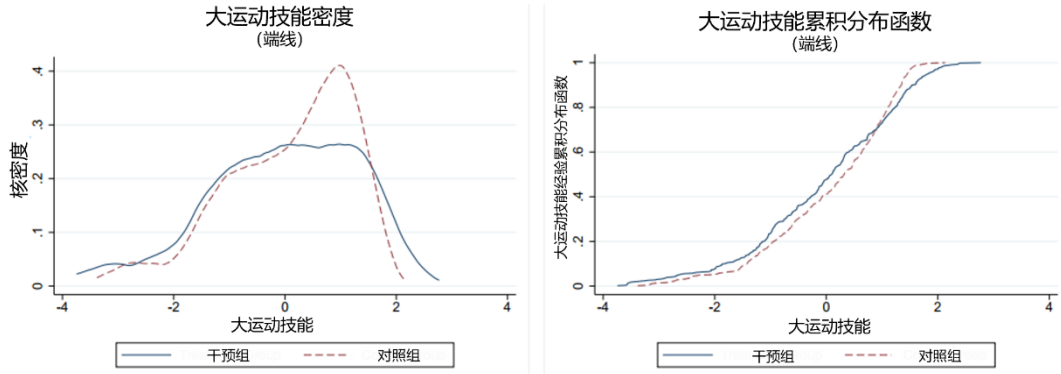
图 6：社交情感技能分布



(1) (2)

图 7：精细运动技能分布

对于大运动技能，几乎没有证据揭示干预效应。对照组和干预组之间的因子分布非常相似。从图 8a 和 8b 可看出，两个大运动技能分布的密度和累积分布函数比较接近。



(1) (2)

图 8：大运动技能分布

总之，该项目极大提高了语言和认知技能、社交情感技能和精细运动技能。但是，认知技能控制分布中的收益并不均匀。这种收益在社交情感技能和精细运动技能方面是一致的。仅从平均干预效应来看，我们发现干预结束时仅在语言和认知技能方面有显著改善，而精细运动技能和社交情感技能没有获得明显提升。通过检测对照组分布的变化，我们能更深入了解哪个人的哪项技能获得了提升。附录 K 给出了估计分布的随机优势检验的扩展阵列。

4 分解 ATE

我们使用自己对潜在技能的估算来了解实验 ATE（平均干预效应）的来源。我们比较了实验干预效应与模型干预效应。

4.1 干预效应的来源

实验产生的平均干预效应可能来自于从技能到任务表现的映射变化，也可能来自于技能的变化。我们研究了每个来源中的数量重要性。

对于 Denver 测验技能 k 中的每个项目 j，其潜在结果为：

$$\tilde{Y}_i^{jk} = X_i'\left[\beta^{j_k,1}D_i + \beta^{j_k,0}(1-D_i)\right]$$

$$+D_i(\theta_i^1)'\alpha^{j_k,1} + (1-D_i)(\theta_i^0)'\alpha^{j_k,0} + \varepsilon_i^{jk}$$

由于恢复了个人潜在技能θ_i^d，我们便能把它们用作等式(3)估计值的输入，以便模拟 Denver 测验评分的平均干预效应。这样得到的平均干预效应点估计非常一致。

表 11：平均干预效应点估计比较

Denver 任务	OLS 模型	因子模型	p 值
	ATE	ATE	
语言和认知	1.113	1.115	0.504
	[0.723, 1.510]	[0.765, 1.454]	
社交情感	-0.115	-0.081	0.556
	[-0.491, 0.275]	[-0.315, 0.152]	
精细运动	0.645	0.569	0.413
	[0.139, 1.158]	[0.136, 0.990]	
大运动	0.219	0.190	0.460
	[-0.294, 0.775]	[-0.071, 0.450]	
$\chi^2(4) = 0.116$			0.998

注：1.括号内 95%置信区间是在村一级使用原始聚类自助法进行的估算。

2.该表中报告的 ATE 估计值以干预前协变量为条件，与表 2 中的第(5)列一致。

3.我们利用沃尔德检验来检查这两种方法得出的 ATE 估计值是否相同。χ^2 检验的 p 值表明，我们不能拒绝这两种方法产生相同 ATE 估计值的假设。

4.2 分解干预效应

实验干预效应不仅可能来自潜在技能θ_i^d的增强，也可能来自从技能到任务表现$\alpha^{j_k,d}$和$\beta^{j_k,d}$的映射关系的变化。为了解入户辅导干预干预效应的来源，我们将项目层面的干预效应分解为两个组成部分：技能到任务映射变化的效果和干预对技能的影响。

对于每一项j_k，实验结果γ_i^{jk}为：

$$\gamma^{jk}(d) = 1(X_i'\beta^{jk,d} + \delta^{jk} + \left(\theta_i^d\right)'\alpha^{j_k,d} + \varepsilon_i^{jk} \geq 0) \tag{5}$$

其中我们假设$\varepsilon_i^{jk} \sim N(0,1)$。入户辅导干预效应来自三个渠道：可观察系数$\beta^{jk,d}$的变化、潜在技能因子$\left(\theta_i^d\right)$的变化以及技能因子载荷的变化。将 $F^1(\theta^1, X)$和 $F^0(\theta^0, X)$分别定义为干预群体和对照群体中(θ^1, X)和(θ^0, X)的分布。对项目 j_k的总体干预效应可分解如下：

$$\begin{aligned}
&\Pr\left(\gamma^{jk,1}=1\right)-\Pr\left(\gamma^{jk,0}=0\right)\\
&=\underbrace{\int\left\{\Phi\left(\left[X'\beta^{jk,1}+\delta^{jk}+(\theta^1)'\alpha^{j_k,1}\right]\right)-\Phi\left(\left[X'\beta^{jk,0}+\delta^{jk}+(\theta^1)'\alpha^{j_k,0}\right]\right)\right\}dF^1(\theta^1,X)}_{\text{来自估计系数 X}}\\
&+\underbrace{\int\left\{\Phi\left(\left[X'\beta^{jk,0}+\delta^{jk}+(\theta^1)'\alpha^{j_k,1}\right]\right)-\Phi\left(\left[X'\beta^{jk,0}+\delta^{jk}+(\theta^1)'\alpha^{j_k,0}\right]\right)\right\}dF^1(\theta^1,X)}_{\text{来自潜在技能载荷}}\\
&+\underbrace{\int\Phi\left(\left[X'\beta^{jk,0}+\delta^{jk}+(\theta^1)'\alpha^{j_k,0}\right]\right)dF^1(\theta^1,X)-\int\Phi\left(\left[X'\beta^{jk,0}+\delta^{jk}+(\theta^0)'\alpha^{j_k,0}\right]\right)dF^0(\theta^0,X)}_{\text{来自潜在技能因子}}.
\end{aligned} \tag{6}$$

请注意，当对照组和干预组中的因子具有相似的可观察协变量分布时，

等式(6)对 X 具有共同的支持，这在我们的样本中基本上得到了满足。[35]表 12 给出了干预效应的分解情况。干预效应的主要驱动力是潜在技能提高。我们在表 6 中表明，干预组和对照组之间的 β 值没有明显差异。因此，β 对干预效应的贡献微不足道。实验引起 α 变化的贡献尚未得到精确估算。我们由此得出结论：干预的主要作用是对潜在技能产生影响。

表 12：干预效应的来源

任务	总净干预效应	来自可观察协变量	来自技能载荷 α	来自潜在技能 θ
语言和认知	1.096	-0.032	0.217	0.911
	(0.184)	(0.189)	(0.192)	(0.187)
		-3%	20%	83%
社交情感	0.258	-0.001	0.049	0.211
	(0.082)	(0.086)	(0.088)	(0.084)
		-1%	19%	82%
精细运动	0.303	-0.009	-0.003	0.315
	(0.085)	(0.088)	(0.189)	(0.315)
		-3%	-1%	104%
大运动	0.150	-0.028	0.062	0.117
	(0.098)	(0.105)	(0.109)	(0.102)
		-19%	41%	78%

注：1.对技能 k 的总体干预效应是$\frac{1}{N_{J_k}}\sum_{j_k}^{N_{J_k}}\left(\frac{\sum_{i=1}^{N_I}\gamma^{j_k,i}D_i}{\sum_{i=1}^{N_I}D_i}-\frac{\sum_{i=1}^{N_I}\gamma^{j_k,i}(1-D_i)}{\sum_{i=1}^{N_I}(1-D_i)}\right)$，假设两个分母均非零，NI 是观测数。

2.为确保可观察协变量在干预组和对照组之间达到平衡，我们考虑了 46 个月以下和 12 个月以上的儿童样本。

3.标准误差记录在括号内。

[35]为了在我们数据中获得对照组和干预组之间的可比样本，我们将样本限制为年龄在 12 个月以上、46 个月以下的儿童。在附录 L 中，我们给出了干预组和对照组之间的年龄分布。

4.3 对潜在技能的干预效应取决于看护者状况

在本部分中，我们根据儿童看护者状况比较了干预效应。我们样本中大约 30-40%是留守儿童，他们存在三种情况：仅父亲外出务工、仅母亲外出务工、父母均外出务工。表 13 提供了对潜在技能因子 θ_i 的干预效应。由于潜在技能因子消除了任务难度级别的影响，因此不同组间的数值更具可比性。从表 13 可看出，干预对弱势儿童（母亲因外出务工不在身边）的干预效应最强。Heckman 和 Zhou（2021 年）指出，当母亲不在身边时，大都是受教育程度较低的祖母充当看护者。

表 13：对潜在技能 θ_i 的干预效应

标准化	(1) 非留守儿童	(2) 留守儿童	(3)	(4)
		母亲外出务工	父亲外出务工	父母均外出务工
		中线		
语言和认知	0.503***	0.730**	0.308*	0.671*
	[0.258, 0.751]	[0.192, 1.330]	[-0.042, 0.661]	[0.049, 1.345]
精细运动	0.463***	0.555	0.669***	0.612
	[0.133, 0.797]	[-0.143, 1.246]	[0.225, 1.130]	[-0.143, 1.391]
社交情感	0.453**	0.825	0.620**	0.622
	[0.075, 0.813]	[-0.174, 1.855]	[0.103, 1.156]	[-0.437, 1.596]
大运动	-0.274**	-0.024	-0.292	-0.074
	[-0.494, -0.050]	[-0.581, 0.472]	[-0.692, 0.080]	[-0.681, 0.462]
		端线		
语言和认知	0.539***	1.443***	0.828***	1.279**
	[0.125, 0.941]	[0.737, 2.255]	[0.456, 1.186]	[0.481, 2.150]
精细运动	0.619***	1.122***	0.831***	1.106***
	[0.428, 0.808]	[0.721, 1.499]	[0.477, 1.166]	[0.662, 1.519]
社交情感	0.245*	0.311	0.560***	0.006
	[-0.013, 0.518]	[-0.283, 1.016]	[0.267, 0.867]	[-0.570, 0.649]
大运动	0.114	-0.514	-0.320*	-0.448
	[-0.105, 0.339]	[-1.207, 0.104]	[-0.649, 0.008]	[-1.187, 0.247]
干预前协变量	是	是	是	是
IPW	是	是	是	是

注：1.括号内 95%置信区间是在村一级使用原始聚类自助法进行估算。

2.标准化评分的均值和方差是根据对照组儿童的汇总样本估算。

3.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5 “慧育中国”干预效应与原牙买加“Reach Up and Learn”项目的比较

从表 14 可看出，就早期可比较结果测量工具而言，“慧育中国”与牙买加“Reach Up and Learn”项目的发展方向一致，而后者已被证明可使人终生受益（Grantham-McGregor 和 Smith，2016 年；Gertler、Heckman、Pinto、Zanolini，Vermeersch、Walker、Chang 和 Grantham-McGregor，2014 年）。我们不能拒绝这两种干预效应相同的假设。“慧育中国”项目若能继续推进，应该也能像牙买加项目一样获得成功。

表 14：“慧育中国”和牙买加“Reach Up and Learn”的干预效应

	“慧育中国”潜在技能因子			
	（干预 21 个月后）			
	社交情感	精细运动	语言和认知	大运动
干预组	0.40***	0 73***	0.75***	-0.10
	[0.21, 0.58]	[0.55, 0.90]	[0.46, 1.05]	[-0.28, 0.09]
	牙买加 Griffiths 测验			
	（干预 24 个月后）			
	表现	精细运动	听力和语言	大运动
干预组	0.63***	0.67***	0.50***	0 34***
	[0.30, 0.95]	[0.34, 1.00]	[0.15, 0.84]	[0.01, 0.67]
p 值	0.35	0.78	0.39	0.15

注：1.“慧育中国”项目中，括号内 95%置信区间是在村一级使用原始聚类自助法进行的估算。

2.牙买加“Reach Up and Learn”项目中，括号内为 95%置信区间。

3.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.最后一行的 p 值是对不同项目的干预效果相等的原假设进行显著性检验。

6 总结

本文评估了大规模儿童早期入户辅导干预项目（即“慧育中国”，该项目借鉴了最初在牙买加开展、广受效仿且成功实施的“Reach Up and Learn”项目）对儿童技能产生的效果。中国国家政策制定充分应用数据，因此此项目的严密证据将对国家政策讨论产生重大影响。

我们评估了儿童自身的潜在技能以及他们会受到该项目的哪些影响。我们开发了一套框架来理解会对儿童技能发展产生干预效应的机制，该机制可根据评估项目的各种任务难度来做调整。这个项目显著提高了儿童的认知和语言技能、精细运动技能和社交情感技能，但其影响程度在基线技能水平上并不一致。我们发现，受影响最大的是最脆弱儿童。潜在技能的提高解释了绝大多数预计的干预效应。我们检验并拒绝了在技能形成经济学中广泛使用的“专用因子”测量模型。测量项目评分取决于多种技能。我们提供了采用各种不同的结果测量工具，并根据任务难度进行调整，来测量潜在技能的分析雏形。使用这些分析方法，我们研究了技能干预对基线儿童技能分布的影响。

参考文献

ANDERSON, T. W., AND H. RUBIN (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, vol. 5, pp. 111–150, Berkeley, CA. University of California Press.

BAI, Y. (2019): "Optimality of Matched-Pair Designs in Randomized Controlled Trials," Unpublished manuscript, University of Chicago.

CAMERON, A. C., J. B. GELBACH, AND D. L. MILLER (2008): "Bootstrap-based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90(3), 414–427.

CANAY, I. A., A. SANTOS, AND A. M. SHAIKH (2019): "The Wild Bootstrap with a "Small" Number of "Large" Clusters," *Review of Economics and Statistics*, pp. 1–45.

CHEN, M., I. FERNÁNDEZ-VAL, AND M. WEIDNER (2021): "Nonlinear Factor Models for Network and Panel Data," *Journal of Econometrics*, 220(2), 296–324.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78(3), 883–931.

ELANGO, S., J. L. GARCÍA, J. J. HECKMAN, AND A. HOJMAN (2016): "Early Childhood Education," in *Economics of Means-Tested Transfer Programs in the United States*, ed. by R. A. Moffitt, vol. 2, chap. 4, pp. 235–297. University of

Chicago Press, Chicago.

GARCÍA, J. L., J. J. HECKMAN, AND A. L. ZIFF (2018): "Gender Differences in the Benefits of an Influential Early Childhood Program," *European Economics Review*, 109, 9–22.

GERTLER, P., J. J. HECKMAN, R. PINTO, A. ZANOLINI, C. VERMEERSCH, S. WALKER, S. CHANG, AND S. M. GRANTHAM-MCGREGOR (2014): "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica," *Science,* 344(6187), 998–1001.

GRANTHAM-MCGREGOR, S., AND J. A. SMITH (2016): "Extending The Jamaican Early Childhood Development Intervention," *Journal of Applied Research on Children: Informing Policy for Children at Risk*, 7(2).

HECKMAN, J. J., AND G. KARAPAKULA (2019): "Intergenerational and Intragenerational Externalities of the Perry Preschool Project," NBER Working Paper 25889.

HECKMAN, J. J., AND J. ZHOU (2021): "Interactions as Investments: The Micro-dynamics and Measurement of Early Childhood Learning," Unpublished Paper, University of Chicago.

HOMVEE (2020): "Early Childhood Home Visiting Models: Reviewing Evidence of Effectiveness, 2011-2020," OPRE Report 2020-126.

HOWARD, K. S., AND J. BROOKS-GUNN (2009): "The Role of Home-Visiting Programs in Preventing Child Abuse and Neglect," *The Future of Children*, 19(2),

119–146.

LIZZERI, A., AND M. SINISCALCHI (2008): "Parental Guidance and Supervised Learning," *Quarterly Journal of Economics*, 123(3), 1161–1195.

LU, B., R. GREEVY, X. XU, AND C. BECK (2011): "Optimal Nonbipartite Matching and Its Statistical Applications," *American Statistics*, 65(1), 21–30.

MAASOUMI, E., AND L. WANG (2019): "The Gender Gap between Earnings Distributions," *Journal of Political Economy,* 127(5), 2438–2504.

RYU, S. H., AND Y.-J. SIM (2019): "The Validity and Reliability of DDST II and Bayley III in Children with Language Development Delay," *Neurology Asia*, 24(4), 355–361.

TSIATIS, A. (2006): *Semiparametric Theory and Missing Data*. New York: Springer.

VAN DER LINDEN, W. J. (2016): *Handbook of Item Response Theory: Volume 1: Models.* CRC Press.

WANG, F. (2020): "Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions," *Journal of Econometrics*.