



IPSOS VIEWS

產品測試新時代： 合成資料的應用潛力

Humanizing AI series,
part two

Colin Ho, Ph.D
Dr. Nikolai Reynolds



在益普索，我們致力於融合人類智慧 (HI, Human Intelligence) 與人工智慧 (AI, Artificial Intelligence)，以推動創新，並為客戶提供以人為本的深刻洞察。

我們的人類智慧來自於在提示工程和數據科學方面的專業知識，以及我們獨特且高品質的資料集。這些資料集將創造力、好奇心、倫理和嚴謹性融入由益普索 Facto Gen AI 平台驅動的 AI 解決方案中，為客戶提供更安全、更快速且深植於人類背景的洞察。

#IpsosHiAi



益普索堅信合成數據
將為市場研究開創嶄
新局面，尤其在產品
測試領域潛力無窮。



合成數據正準備顛覆全世界。無論是在醫療領域加速新藥研發、在金融業模擬詐欺交易情境，或是推動汽車產業的自駕車測試，合成數據已在各個商業應用場景中證明其價值。

益普索堅信合成數據將為市場研究開創嶄新局面，尤其在產品測試領域潛力無窮。不過，許多企業對於合成數據的品質或評估方式仍存有疑慮。本報告正是為了填補這些認知落差而生。

身為全球規模最大、地位領先的產品測試顧問公司，益普索始終走在運用前沿技術協助全球企業加速創新與成長的最前線。

益普索在合成數據產品測試方面的深度見解，為讀者提供：

- 生成及評估高品質合成資料集的實用建議
- 消費性商品與服務產品測試的具體應用案例

目前市場研究業界存在多種不同類型的合成數據，各有其優劣之處。本報告將聚焦於數據強化技術，也就是運用合成數據來豐富既有的資料集。

合成數據應用方法

量化數據增強

數據提升 Data Boosting

透過合成數據豐富既有資料集，增加觀測樣本數量

數據插補 Data Imputation

運用機器學習技術填補遺漏的資訊

資料融合 Data Fusion

透過機器學習技術結合多個資料集，產出全新的資料集

虛擬受眾模擬研究

虛擬角色機器人 Persona Bots

專為模擬特定消費族群而設計的客製化數位助理

合成樣本群組 Synthetic Panels

運用完全由合成受訪者構成的虛擬樣本

我們首先從如何生成高品質合成數據以及品質評估方法的整體概念談起。由於合成數據的實際效果與細節差異，必須在特定應用場景中才能被充分理解，因此我們將深入探討合成數據在產品測試領域的具體應用。

本報告僅聚焦於合成數值資料的產生過程，這也是量化研究人員最常採用的資料格式。至於合成影像、影片、數據插補或虛擬角色機器人等應用領域，雖同樣屬於合成數據的大範疇，但不在本文討論範圍內¹。

正如益普索白皮書 Synthetic Data: From Hype to Reality¹ 中所闡述，合成數據是透過訓練模型來模擬真實世界資料的統計特徵與模式，進而產生的人工資料。



如果 AI 沒有使用與您企業相關的真實世界資料進行訓練，就無法產生具備真實資料特性的合成數據。道理就是這麼簡單！



合成數據的生成與評估

我們運用數據在現實世界中做出更明智的商業決策，而合成數據可透過多種方式來輔助決策制定。因此，儘管合成數據並非對應真實事件或真實人物，但仍須模擬真實世界資料的統計特徵與模式。這衍生出兩個核心問題：

- 01 要生成高度貼近真實資料的合成數據，需要具備哪些條件？
- 02 如何評估合成數據與真實世界資料的相似程度？

在 AI 能夠生成反映真實世界資料的合成數據之前，必須先使用真實世界的資料來訓練 AI。正如益普索 Humanizing AI 系列首篇白皮書中所述，AI 本質上只是演算法，在接受訓練前並不具備智慧。

它必須透過學習真實世界的資料，才能掌握我們所認知的「智慧」。這正是本報告最重要的核心觀念：如果 AI 沒有使用與您企業相關的真實世界資料進行訓練，就無法產生具備真實資料特性的合成數據。道理就是這麼簡單！

而評估合成數據的品質，方法同樣直觀。合成數值資料至少應在常見統計指標上反映真實世界資料的特徵——如平均值、資料分布、變異數，以及變數間的關聯性（例如相關係數）。透過這些常見指標直接比較合成資料與真實人類的資料，可讓我們了解兩者的相似程度。合成數據越接近真實人類資料，使用時的風險就越低，但始終存在一定風險，因為合成數據永遠無法在所有層面完美複製真實資料。因此，我們應該只在願意承擔一定風險的前提下才使用合成數據。

運用大型語言模型 (LLMs) 生成合成數據

合成數據的生成方法可分為兩大類：LLMs (Large Language Models, 大型語言模型) 與非 LLMs，兩者分別以文本處理與數值運算為基礎來區分。目前市面上常見或開放的 LLMs，已透過大量公開資料進行預先訓練，例如網站內容、電子書、社群媒體貼文等，因此能夠針對這些領域產出品質不錯的合成數據。

然而，若要生成貼近真實情境的資料，現成 LLMs 仍有不少限制 (見圖 1)^{2,3}。首先，許多較日常或具高度私密性的主題，難以透過網路取得公開資料。其次，由於訓練資料高度集中於西方英語系國家，使得 LLMs 往往存在地區性的偏誤。研究顯示，LLMs 所生成的內容在

文化觀點上，與英語系及新教歐洲國家的價值觀較為相近，與其他國家則存在落差⁴。第三，模型所使用的資料往往滯後於現況，資訊更新速度也不夠即時。

因此，要運用 LLMs 生成高品質合成數據，關鍵在於使用與研究主題高度相關、具時效性，且反映在地情境的真實世界資料來重新訓練模型。這不僅需要掌握統計與資料科學知識，還得投入大量時間蒐集、整理與評估資料，才能確保模型輸出的數據真正貼近真實世界的統計特徵與脈絡⁵。

圖 1：人類資料集 vs. 人類 + 合成資料集之比較



來源：益普索

運用非 LLM 方法生成合成數據

在 LLMs (大型語言模型) 成為焦點之前，資料科學家早已開始運用深度學習 (Deep Learning, DL)⁶ 來生成合成數值資料⁷。這類深度學習技術，包括現今應用於 LLMs 的架構，在生成合成數據方面皆展現強大實力，但各自擁有不同的優勢與適用情境。

LLMs 尤其擅長產生貼近人類語言的文本內容，能創造出結構完整、語意豐富且具情境脈絡的文字，因此在內容生成、語言翻譯、聊天機器人等應用場景中具備極高的實用價值⁸。

非語言模型的 DL 技術，正在為數值資料的合成開創新局。

相較於更擅長語言任務的 LLMs (如 ChatGPT)，DL 更擅長針對數值資料進行建模，能夠生成在統計結構上逼近真實資料的合成數據。透過針對特定領域或市場場景進行訓練，DL 模型也能提供更具彈性的客製化合成方案。在益普索，我們長期致力於運用 DL 技術進行數據合成。接下來，我們將細說明如何運用 DL 模型來產出產品測試用的合成資料，並根據不同產品逐一分析。

為何選擇產品測試作為切入點？

雖然在市場研究以外的領域，合成數據的許多應用主要著重於匿名性 (如醫療數據匿名化以保護隱私)，但在市場研究中，多數企業尋求的核心效益是節省收集真實世界資料的成本與時間。

但隨著線上數據蒐集日益便捷且成本降低，我們必須審慎權衡，合成資料所帶來的時間與金錢效益，是否足以彌補其準確度上的潛在損失。舉例來說，透過益普索旗下的敏捷測試平台 Ipsos.Digital，一位美國研究員僅需約 60,000 元 (2,000 美元)，便能在 24 小時內完成 300 份問卷調查。如果改用合成資料模擬這些人可

能給出的回答，時間或許能縮短至 12 小時，成本也降為 45,000 元 (1,500 美元)——但準確度也會隨之降低。在這樣的情境下，與其為了節省 15,000 元 (500 美元) 和 12 小時而犧牲數據品質，許多研究人員可能會選擇取得真實受訪者的回饋。畢竟，多等一點、多花一些，就能換來更可靠的結果！

正因為合成資料存在這種固有的利弊權衡，我們希望在研究成本普遍較高的情境中測試其應用效果。而產品測試完全符合這個條件，因為它往往涉及到多項成本考量：



製造成本：從產品製作、採購原型及產品，到為了測試需求而進行產品外觀遮蔽的費用



物流成本：包括產品寄送、退貨處理，甚至是空包裝的銷毀成本



樣本招募成本：受訪者招募費用，尤其當企業需要精準篩選特定用戶族群時

由於製造、物流和樣本招募的龐大開支，若能減少參與測試的人數，就能有效降低整體支出。這並不代表合成資料無法應用於市場研究的其他領域，而是說明當成本節省效益有限時，使用合成資料的風險可能得不償失。

人類的參與仍不可或缺

產品體驗本質上無法被演算法完全複製。AI 無法單獨感知人類在使用產品時的五感體驗、情緒反應、心理期待，以及使用情境所帶來的影響。因此，我們應用合成資料的目標，不是要取代人，而是讓人的洞察更加精準。

我們面臨的挑戰在於，若希望導入合成資料來協助產品測試，我們至少還需要多少名真實受訪者，才能確保研究結果的可靠性？為達成此目標，益普索的 Innovation 團隊展開了兩項研究：

第一階段研究

找出在不使用合成資料的情況下，最少需要多少真實受訪者樣本，才能達到近似大樣本 (如 200-300 人) 的產品測試結果

在**第一階段研究**中，益普索團隊從產品測試資料集中擷取樣本，參考了來自 4 萬名受訪者、涵蓋 185 項消費性包裝商品 (CPG) 的測試數據⁹，目的是評估最少需要多少真實受訪者樣本，才能與大樣本的產品測試結果達到足夠高的相關性。研究

第二階段研究

測試這個小樣本若搭配合成資料，是否能重現全部使用真實受訪者樣本的測試結論

顯示，當最受歡迎與最不受青睞的產品在表現上至少拉開 8% 的差距時，只需要 50 名真實受訪者，就足以正確地還原出前後名次 (相關係數 $r = 0.8$)。

之所以能用如此少的參與者數量就能得到近似大樣本的測試結果，原因在於產品測試資料的差異主要來自產品技術層面



產品體驗本質上無法被演算法完全複製。AI 無法單獨感知人類在使用產品時的五感體驗、情緒反應、心理期待，以及使用情境所帶來的影響。



的變動 (例如糖分含量多寡)。因此，產品測試中觀察到的差異通常遠小於其他研究領域，如消費者態度或品牌認知評估等。

理論上，企業應該根據這些研究發現，採用 50 人的樣本進行產品測試，特別是在風險相對低的早期測試階段。

然而，大多數企業並未採用這種做法，主要有兩個原因：

- 01 50 人的樣本無法提供足夠的細分洞察。有時企業需要分析消費者人口中特定的群體。
- 02 50 人的樣本規模在統計上不足以辨別差異。對依賴「統計顯著性」作為決策門檻的企業來說，這恐怕無法作為行動依據。更重

要的是，樣本過少會增加「第二型錯誤 (Type II Error)」的風險——也就是產品確實表現出差異，但卻無法被偵測出來。

因此，我們提出另一種折衷方式：與其全面招募 200 人，不如先找 50 名「真實消費者」參與測試，再利用他們的反應資料訓練 AI 模型，產出另外 150 位具相同特徵與行為反應的合成受訪者 (不重複或重新抽樣原本的 50 人)。可以將這 50 名人類視為訓練 AI 的「種子」樣本，讓 AI 能夠準確生成模擬人類對產品反應的合成資料。

這就是**第二階段研究**的目的：驗證在加入合成資料後，小規模的人類樣本是否仍能得出與全數使用真實受訪者時相符的研究結果。為提升研究結果的普遍通用性，我們在本次合成資料測試中納入了不同國家與產品類別的樣本。同時，我們也

曾測試較大的種子樣本規模 (如 75 人與 100 人)，但為簡明呈現，本報告僅分享測試 50 人樣本的研究成果。

接下來，研究團隊使用這 50 位真實受訪者的資料來訓練 DL 模型，以生成合成資料。我們選擇不使用現成的大型語言模型 (LLM)，主要原因是 LLM 預先訓練的公開資料並未涵蓋人們對特定類別產品的多感官體驗 (如產品原型、新配方)，也缺乏

具實證價值的受訪者數據。此外，我們也沒有考慮採用資料加權的方式來取代深度學習，因為加權處理無法為子群體分析增加額外的樣本數量，而且在產品測試領域通常難以被接受為有效依據。



我們將「全部使用真實受訪者的樣本」與「合成人工資料輔助樣本」的結果進行比對驗證。

整體而言，合成數據確實有效

我們將「全部使用真實受訪者的樣本」與「合成人工資料輔助樣本」的結果進行比對驗證 (見圖 1)¹⁰。需要強調的是，我們的方法並非單純複製貼上既有的受訪者數據。

實驗結果顯示，兩組資料在多方面都高度一致，包括：

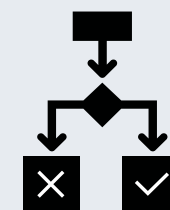
- 產品的相對表現 (如排名、是否達到統計上的顯著差異)

- 數據分佈 (如受訪者在各題目選項上的回應分佈)
- 各變數之間的關聯 (例如整體滿意度與產品特性之間的相關性)

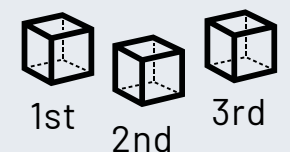
最關鍵的是，儘管兩個資料集在變異數上存在差異，但在我們測試的所有資料集中，都得出了相同的商業決策結論 (見圖 2)。

圖 2：判斷「合成人工資料輔助樣本」能否準確重現「全部使用真實受訪者的樣本」結果的評估標準

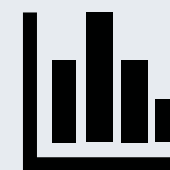
商業決策與行動依據



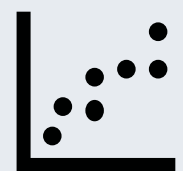
產品排名與差異顯著性



數據分佈



相關性



來源：益普索

使用合成資料進行產品測試的一大優勢，在於可以有效擴充樣本數，尤其針對難以招募的族群。一旦樣本數提升，原本不具統計顯著性的差異，可能因此浮現出來。例如在一項測試中，我們針對特定品牌的產品使用者生成了合成回應資料。

在全部使用真實受訪者的樣本中，每項產品大約有 100 位品牌使用者。儘管三款產品之間出現一些差異，但並未達統計顯著性。然而，當我們針對每款產品額外加入 100 位合成品牌使用者後，產品之間的差異因樣本規模擴大而達到統計顯著性 (見圖 3)。

圖 3：以合成品牌使用者擴增樣本



來源：益普索



當我們需要特定目標族群對產品的意見回饋時，必須根據商業目標設定明確的招募配額。

使用合成資料仍須謹慎

儘管我們展示了合成資料的潛力，但正如先前所提，它終究無法完全取代真實世界的資料。使用「部分真實、部分合成」的混合資料，是否能準確重現全部使用真實樣本的研究結果，關鍵在於以下幾點：

• 真實樣本在 AI 訓練中的代表性。



在我們的前導測試中，若未將 50 位真人受訪者的樣本結構調整為與原始 200 人樣本一致，而直接用這 50 位的資料來生成合成數據，結果往往會出現偏差。實驗顯示，在這樣的情況下，大約有兩到三成的樣本會無法準確反映純真人樣本中各產品的表現。

換句話說，當真人樣本缺乏代表性時，生成的合成資料可能會失真，導致錯誤的分析結果。不過，一旦我們讓這 50 位受訪者的組成與原始 200 人樣本一致，再透過 AI 生成額外的 150 筆合成資料，最終所做出的商業決策將與使用完整 200 位真人數據的結果完全一致。總而言之，當我們需要特定目標族群對產品的意見回饋時，必須根據商業目標設定明確的招募配額。就如過去在大規模抽樣中所做的一樣，這確保我們能夠抽樣到正確的目標族群。

• 產品本身的差異化程度也會影響結果。



如果純真人樣本中的產品表現差異明顯，AI 便能夠輕易識別這些差異並在合成數據中重現這些結果。但若產品彼此之間並無明顯統計差異，AI 就如同純真人樣本一樣，都難以做出有效區分（見圖 4）。在我們於不同國家與產品類別中執行的前導測試中，我們發現擴增樣本 (boosted samples) 的另一項優勢，特別是在產品之間差異不明顯時尤其有幫助。

當我們將原始樣本細分為子群體（例如現有品牌使用者與非品牌使用者）後進行分析，透過 AI 生成的合成資料來擴增子群體的樣本數，能有效提升對產品表現的差異。這讓我們在子群體中能夠偵測到更多由目標族群偏好所驅動的結果，也揭示出原本可能被忽略的產品偏好。

圖 4：合成資料是否準確，關鍵在於產品本身的差異性



最後仍需提醒，本研究的驗證結果僅限於「產品測試」，並不代表所有市場研究都能以 50 位受訪者完成。讓產品測試能夠僅使用 50 人的條件，在其他研究領域未必存在。作為全球領先的產品測試顧問公司，益普索從自身的資料庫與實務經驗中汲取相關經驗。

此外，產品測試具備一些有利於合成資料品質的特性，例如涉及人類感官回饋的產品屬性。因此，每一種合成數據的應用，仍需根據特定目的（如市場區隔）進行個別驗證。



資料向來被視為企業的命脈，支撐決策並驅動成長。



合成資料的價值：從炒作走向現實

資料向來被視為企業的命脈，支撐決策並驅動成長。因此，合成資料能即時且大量生成的潛力，自然備受關注。然而，公眾對合成資料的態度仍相當兩極。有些提供合成資料服務的公司可能會宣稱：「這就是你所需要的一切，不再需要真人參與！」而一些較為謹慎的研究人員則採取觀望態度，暫不使用這項技術。這些對立觀點，其實反映了一種二元化的思維：究竟是好還是壞？應該選擇合成資料，還是真實資料？我們的初步研究指出，這並不是非此即彼的問題——如同其他研究工具一樣，合成資料也有其優勢與限制，且適用於特定情境。

在這項前導研究中，我們證實了 AI 能夠生成合成數據來模擬真實世界的數據，但前提是必須先以優質的人類數據進行訓練。

因此，重點不是選擇其一，而是需要將合成資料與真實資料搭配運用。合成資料的準確性不是絕對的好或壞，而是取決於應用情境。例如，當產品在真實人群中的差異極小，我們可能需要觀察特定子群體。因此，若決定採用合成資料，就必須理解並接受其在某些條件下可能不奏效的風險。身為研究人員，我們的責任是確保只在能夠最大化成功機率的條件下使用合成數據。相較於使用較小的樣本數，運用合成數據有諸多優勢，包括能夠進行子群組分析、維持統計效力，以及執行更複雜的分析。

整體而言，我們不認為合成數據應該完全取代人類——至少在產品測試領域不應該如此。在 1997 年的電影《心靈捕手》中，已故演員羅賓·威廉斯 (Robin Williams) 飾演一位教授，指導由麥特·戴蒙 (Matt Damon) 飾演的天才青年。這名青年擁有過人的記憶力，能從書本中吸收大量知識。在其中一幕，教授對他說：「但我敢說你不知道西斯汀禮拜堂的氣味。你從來沒有真正站在那裡，仰望那幅美麗的天花板。」這句話點出一個核心觀念：真正的理解源自於生活經驗，而非從書本、圖片或影片中模擬而來的知識。

就像電影中的天才一樣，AI 可以被灌輸世界上所有存在的知識，但 AI 永遠無法像人類一樣體驗世界。人類的獨特價值，在於我們能感受陽光灑落臉龐的溫度、聆聽旋律中的情緒、讚嘆夕陽的顏色與變化——這些都是獨特而美好的體驗，無論 AI 多麼先進，都永遠無法完全複製。人們對產品或生活的反應，亦不僅止於知識與認知。身體與感官的經驗，同樣深刻影響我們的判斷與選擇。



真正的理解源自於生活經驗，而非從書本、圖片或影片中模擬而來的知識。

關鍵洞察

01

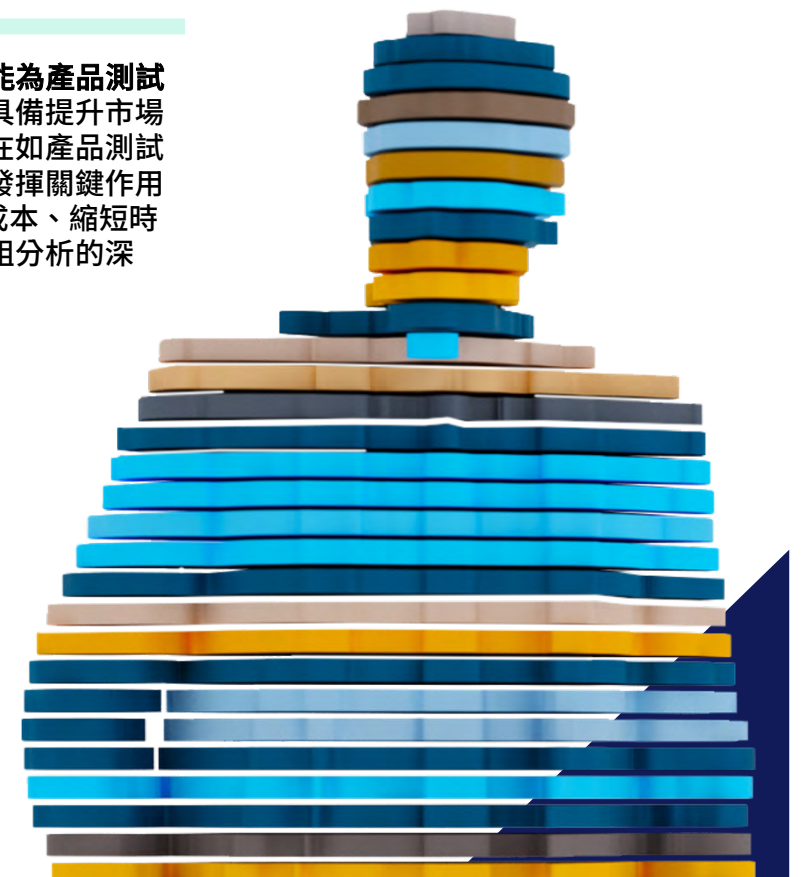
合成數據永遠無法取代人類。
AI 無法單獨重現我們的產品體驗，因為這些體驗結合了五感、情緒、期待與情境脈絡。因此，我們的目標是透過合成數據來強化人類的回饋，而非取而代之。

02

準確性取決於訓練資料的品質。合成數據的價值並非非黑即白 (好與壞)，它的準確性受到多項因素影響，包括我們試圖複製的資料差異程度，以及用來訓練 AI 的真實樣本是否能代表目標族群。因此，使用合成數據應該採取策略性思考，權衡相關的風險與效益。

03

當準確度達標，合成數據能為產品測試創造顯著價值。合成數據具備提升市場研究靈活性的潛力，尤其在如產品測試等資源密集的領域，更能發揮關鍵作用——不僅能有效降低研究成本、縮短時程，還能進一步強化子群組分析的深度。



尾註

1

Guidi, M., Hubert, B., Sava, C., & Timpone, R. (2024). Synthetic Data: From Hype to Reality – a guide to responsible adoption. Ipsos POV

2

Illic, Maya, Bangia, Ajay, Legg Jim (2024). Conversations with AI Part V. Is there depth and empathy with AI twins? Ipsos Views

3

Moore Chris, Stronge Cameron, Bhudiya, Manjula (2024). Judgment Day: The Machines Have Arrived – But how good are they at answering choice experiments? Sawtooth Conference.

4

Yan Tao, Olga Viberg, Ryan S. Baker and René F. Kizilcec (2024). Cultural bias and cultural alignment of large language models. PNAS Nexus, Vol. 3, No. 9

5

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate (2022). Out of One, Many: Using Language Models to Simulate Human Samples. arXiv.

6

AI-based Deep Learning is a way for computers to learn by analyzing large amounts of data and finding patterns, much like how humans learn from experience. It uses neural networks inspired by the human brain to recognize information and make decisions without being explicitly programmed.

7

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio (2014). Generative Adversarial Networks. arXiv.

8

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

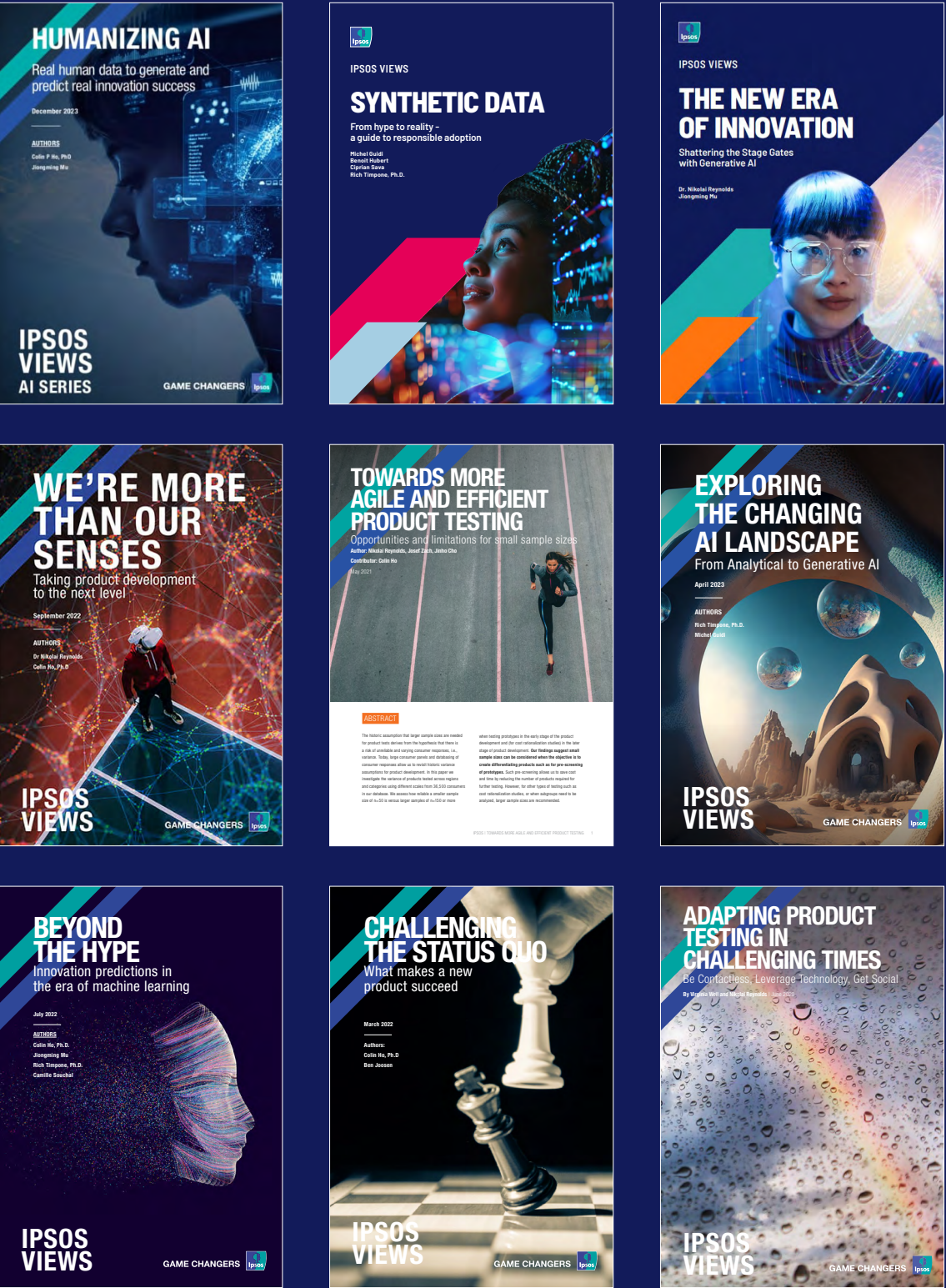
9

Reynolds, N., Zach, J., Cho, J, Ho, C. (2021). Towards more agile and efficient product testing. Opportunities and limitations for smaller sample sizes, Ipsos POV.

10

We also compared the results from the all-synthetic data to the all-human data; the findings are like those reported in this paper

閱讀更多



FEBRUARY 2025

產品測試新時代：合成資料的應用潛力

Humanizing AI series, part two

作者

Colin Ho, Ph.D

Chief Research Officer,
Innovation and Market Strategy
& Understanding, Ipsos

Dr. Nikolai Reynolds

Global Head of Product Testing,
Ipsos

The **IPSOS VIEWS** white papers are produced by the **Ipsos Knowledge Centre**.

www.ipsos.com

@Ipsos

