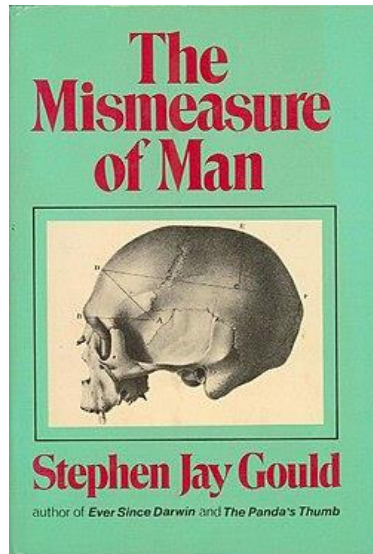

賢愚千載知誰是？

余創豪 chonghoyu@gmail.com



過去，我曾談過精神醫學專家的判斷如果被絕對化，也可能造成一種「專業暴政」。其實其他專門學者亦然，特別是那些試圖以不同「科學方法」去甄別人的智慧、能力和前途的專家。十九世紀，法國人類學家布羅卡（Paul Broca）量度過不少頭蓋骨之後，認為人的智慧與腦袋大小有關。當德國學者測出德國人的頭骨較大時，布羅卡又提出各種身體比例和腦區發達程度的解釋，設法證明法國人依然較具智慧。後來古生物學家古爾德（Stephen Jay Gould）在名著《人類的誤測》（The Mismeasure of Man）中重新檢視這一類研究。他指出，問題並不一定在於科學家偽造數據，而可能是研究者自己也沒有察覺的偏見（unconscious bias）：先有了某種觀念，然後在選擇、分析和解釋數據時，不知不覺採用了有利於自己原來看法的標準。

二十世紀初，美國政府在大規模徵召士兵時，需要以測驗甄別不同人的能力，智力測驗亦因而得到很大的發展，此後各種測量人的智力、能力、性向和潛能的方法更是五花八門。1975年，威爾遜（E. O. Wilson）出版《社會生物學》（Sociobiology），嘗試從演化生物學的角度解釋動物以至人類的一些行為模式。在全書最後討論人類的部分，他提出男女若干行為傾向和傳統性別分工可能具有演化上的基礎。這些說法隨即引起激烈爭論，包括古爾德在內的一些學者批評這類解釋容易陷入「生物決定論」（biological determinism），把既有的社會制度和性別角色披上一件「科學」的外衣。

當時不少研究又指出，一般而言男學生在某些數學測驗上表現較好，女學生則在某些語文能力方面較突出。這些所謂「科學發現」一旦進入社會，便很容易變成另一回事：男孩子被鼓勵去讀數理科，女孩子則被鼓勵讀文科；男生讀文科可能被人取笑，女生讀數理也可能被認為不合適。另一方面，當時也有新的研究開始指出，這些差異未必完全來自先天能力，家庭、學校、教師期待和社會鼓勵都可能產生影響。可惜，把「男」和「理科」、「女」和「文科」聯繫在一起的觀念，在華人社會也並不罕見。有些香港女學生雖然理科成績很好，但老師卻仍然建議她們修讀文科。有些華人女學生修讀工程科，人們便誇讚

是「女強人」，這類說法表面上是稱讚，實際上仍然不知不覺地把學科和某種性別聯繫在一起。當中的潛台詞是：理科是艱辛的科目，一般來說男生去讀理科，一名女性必定是「強人」，才可以讀理科。我甚至曾經聽朋友評論一位作家，說他的文章一定很好，因為他的氣質像白先勇般的「女人型」。

心理學史上還有一個很著名的例子，就是所謂「比馬龍效應」（Pygmalion effect）。1960 年代，哈佛大學的羅森塔爾（Robert Rosenthal）和同事到一所小學進行研究。他們告訴教師，某種測驗可以找出具有很大智力發展潛能的學生；測驗之後，研究者其實只是隨機選出部分學生，卻告訴教師這些孩子具有很高的「智力爆發」潛能。後來研究者回訪，報告部分被標示為具有潛能的學生在測驗表現上有較大的進步。這個研究因而成為「自證預言」（self-fulfilling prophecy）最著名的例子之一：老師如果相信某些學生具有潛能，可能不自覺地給他們更多注意、鼓勵和學習機會；學生得到這些正面訊息後，也可能增加自信，最後真的改善表現。可是，我當時想到另一個問題：如果把實驗倒轉過來呢？假如有人隨意在一班選出十個學生，然後由一個掛着「哈佛大學測驗」名義的專業工具宣布他們是「蠢才」，他們的命運又會怎樣？

現實世界中，類似的事情其實每天都可能發生。當時美國一些大學使用一種我們俗稱「密芝根大學試」的英文測驗去評定外國學生的英語水平；以我個人的經驗，我一直覺得它甚至不及托福可靠。要進入美國研究院，GRE 或 GMAT 又是許多人必須通過的關卡。我當時也十分懷疑，這些考試能否真正反映一個人的學術能力；其中一些艱深英文字彙，甚至連不少美國人日常生活中也很少使用。我並不是說所有測驗都沒有用，而是懷疑我們是否太容易把一個測驗分數，由「一項參考資料」變成對一個人的整體判斷。

我自己便有過非常深刻的經歷。年輕時在香港，我面對填鴨式教育和許多公開考試的壓力，當時的學業表現並不好。一位很有權威的專家曾經對我說：「你不適宜讀書，你最好找一些比較簡單、不需要太多學術訓練的工作。」試想像，如果我當時完全相信了他的專業判斷，認定自己真的「不適宜讀書」，心理學所說的自證預言也許便會出現，只不過那將是一個負面的自證預言。我的人生很可能因此走上完全不同的道路，我也許會放棄繼續求學，更不可能在多年之後坐下來寫這篇文章。

因此，我當年在美國求學時有一個很深的感受：至少在我所接觸的一些學校裏，考試成績並不一定一次便決定一個人的命運。相比之下，我回想自己早年在香港接受教育的經歷，便覺得考試、分流和所謂「能力判斷」對一個人的前途影響實在太大。多少由專家設計、聲稱可以分辨能力的考試，多少基於專業知識而對一個人作出的判斷，如果我們把它們由「參考」提升為不可懷疑的判決，便可能變成一種假科學之名而行的暴政。我並不是

主張破除一切標準；沒有標準，教育和社會同樣難以運作。問題是：標準和級別應該是參考，而不能被絕對化。

如果我們因為一個人學歷高、考試成績好，便認定他一定特別聰明；他自己也逐漸相信自己凡事都比別人高明，那便容易錦上添花。相反，一個人在早年考試失敗，如果旁人從此不再看好他，他自己最後也不再看好自己，那個分數便可能由一次測驗結果，慢慢變成一個人的命運。美國教育當然也有很多問題，但我當年很欣賞其中一點，就是一些研究院在考慮成績之外，也願意讓申請人解釋有甚麼特殊因素影響自己的表現。我記得申請研究院時，表格上有一欄讓申請人說明這類情況；我很認真地寫：「我在美術系得不到甲級，因為美國老師不懂得欣賞我的中國畫。」結果，我還是被錄取了。

1991 年 7 月 10 日

原載《澳門日報》

2026 後記：從智力測驗到人工智能評分

三十五年後重讀這篇文章，我不打算用今天的知識去改寫 1991 年的自己。一篇舊文章如果把後來才出現的研究偷偷放回原文，文章固然會變得「正確」一些，卻不再是真實的歷史。1991 年的我知道甚麼、不知道甚麼，當時心理學界和教育界流行甚麼觀念，本身就是文章的一部分。但是，三十五年來確實有不少研究，使我今天再看這些問題時有了不同的理解。首先是古爾德。我當年讀《人類的誤測》，很受他分析 Broca 等人的方式影響。今天回看，Broca 的頭骨測量和他對種族、性別與智力的若干推論，仍然被科學史研究者視為十九世紀科學如何受到當時社會偏見影響的重要例子；但另一方面，Gould 本人對十九世紀頭骨研究史的一些分析，後來也受到重新檢驗和批評，尤其是他對另一位頭骨研究者 Samuel Morton 的分析。後來研究者重新找到和分析 Morton 的原始資料後，認為 Gould 對其測量偏差的部分指控並不成立，但 Morton 如何解釋數據以及當時的種族觀念仍然是另一回事。這件事給我的啟示反而更加有趣：批評偏見的人，也不因此免於被後人重新檢驗。科學真正值得信任的地方，也許並不是科學家永遠不會犯錯，而是沒有任何一位科學家應該擁有最後的發言權。

我在 1991 年文章中談到男女的數學能力，今天看來也需要更加謹慎。後來的大型統合分析發現，男女平均數學表現的差異通常非常小；例如一項綜合 242 項研究、超過 128 萬人的 2010 年統合分析，整體差異只有 $d = 0.05$ ，研究者認為男女的數學表現基本相近。

跨國研究同時顯示，不同國家的差異並不一致，教育、文化、機會和性別角色等因素都與結果有關。換言之，1991 年我真正擔心的事情仍然成立，但理由應該說得更加準確：最大的危險不是證明「男生和女生完全沒有任何差異」，而是把群體平均數上的某些差異，變成對眼前一個人的命運判決。即使某一群體平均而言真的存在某種統計差異，也不能由此推論面前這一個男孩一定適合理科、這一個女孩一定適合文科。統計上的群體，不是個人的命運。

我當年對 Pygmalion effect 的描述也比較接近當時流行的教科書版本：教師相信孩子是天才，孩子便真的「突飛猛進」。後來幾十年的研究使情況變得複雜得多。Rosenthal 和 Jacobson 的原始研究一直受到方法上的討論，後續研究也不是每一次都能複製出同樣強烈的效果；不過，大量研究仍然支持一個比較溫和、也比較可信的結論：教師的期待確實可以影響教師與學生的互動，而這些互動在某些條件下又會影響學生的自我概念、學習機會和學業表現。所以，1991 年我問的那個問題，我今天仍然會問：「如果老師相信一個孩子沒有能力，會發生甚麼事？」只是今天我不會再把答案說得那麼簡單。

至於 GRE，我當年對它頗不客氣。三十五年後，也應該公平地承認，後來的研究並不支持「GRE 完全沒有用」這種簡單結論。一項涵蓋 82,659 名研究生的 2001 年統合分析發現，GRE 對研究院成績、綜合考試、教師評價等確實具有一定預測能力；可是，2024 年另一項涵蓋 201 項研究的統合分析得到比較保守的結果：多數個別效果並不顯著，而 GRE 對所研究結果的整體解釋量平均約為 3.24%，研究者並發現其預測關係隨年代有所下降。因此我現在反而覺得，這是一個很好的例子：測驗可以有用，而又同時非常有限。問題從來不是「測驗究竟有沒有用」這個非黑即白的選擇，而是我們給它多大的權力。一個分數可以提供資料；它不應該變成一個人的定義。

有趣的是，當自己成為教授之後，起初以為從此不需要或者很少會受到「分數」的困擾，然而，美國很多大學教授的合約和升遷在很大程度上到取決於學生的評分。其實，有很多因素會影響學生的評分高低，並不純粹或者大部份反映了教授的能力，那些教授以外的因素包括了學科本身的難度、學生的水平、上課的時間，甚至是教授本身的樣貌。Nate Jornell、Tarquino Sanchez、Wolfgang Stroebe……等學者的研究已經否認了學生評分對改善教學的作用。不幸地，很多大學都將學生評分絕對化，結果將原本好的老師逼成對學生投其所好，導致了「成績通脹」（grade inflation）。

 Listen

Original Articles

Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis

Wolfgang Stroebe 

Pages 276-294 | Published online: 13 May 2020

 Download citation  <https://doi.org/10.1080/01973533.2020.1756817>  Check for updates

[Full Article](#) [Figures & data](#) [References](#) [Citations](#) [Metrics](#) [Licensing](#) [Reprints & Permissions](#) [PDF | EPUB](#)

Abstract

Student Evaluations of Teaching (SETs) do not measure teaching effectiveness, and their widespread use by university administrators in decisions about faculty hiring, promotions, and merit increases encourages poor teaching and causes grade inflation. Students need to get good grades, and faculty members need to get good SETs. Therefore, SETs empower students to shape faculty behavior. This power can be used to reward lenient-grading instructors who require little work and to punish strict-grading instructors. This article reviews research that shows that students (a) reward teachers who grade leniently with positive SETs, (b) reward easy courses with positive SETs, and (c) choose courses that promise good grades. The study also shows that instructors want (and need) good SETs.

[Previous article](#) [View issue table of contents](#) [Next article](#)

Related research

[People also read](#)

[Student evaluations of teaching: inadequate assessment of faculty performance](#)
Henry A. Hornsby
Cogent Education
Published online: 2020

[Unbiased, reliable, and valid: can still be unfair](#)

而到了 2026 年，這個問題又出現了一個我在 1991 年完全想像不到的新版本：人工智能。三十五年前，我擔心的是由心理學家、教育學家或考試機構設計一個測驗，再用測驗分數替人分類；今天，電腦系統可以閱讀履歷、篩選求職者、分析網上測驗，甚至評估錄影面試，再把一個人變成一個分數、排名或者「是否值得進入下一輪」的判斷。美國平等就業機會委員會（EEOC）已經明確指出，AI 和自動化技術目前可用於招聘、履歷篩選、錄影面試、員工評估、晉升甚至解僱，而這些系統同樣可能產生歧視性結果；紐約市亦已規定某些用於招聘和晉升的自動化就業決策工具，在使用前必須接受偏差審核，並讓求職者知道自動化工具正在參與評估。

這使我想到了 1991 年文章裏那個假想的「哈佛大學測驗」。當年，如果一位穿着白袍的專家拿着一張測驗結果，說「科學證明這個孩子沒有潛能」，一般人很容易相信他；今天，白袍可能不見了，取而代之的是一個演算法，螢幕上只出現「Score: 37」或者「Not recommended」。我們甚至可能更加相信它，因為我們會想：這不是某個人的偏見，而是電腦算出來的。但電腦所使用的資料是人收集的，衡量甚麼是人決定的，甚麼叫「成功」是人定義的，模型從甚麼歷史資料中學習也是人選擇的。NIST 對 AI bias 的研究也提醒我

們，人工智能不但可能承繼原來存在的偏差，而且可以把有害偏差以更大的速度和規模擴散。

於是，1991 年的問題並沒有消失，它只是換了形式：從頭骨大小，到 IQ；從 IQ，到公開考試；從公開考試，到 GRE；再從 GRE，到演算法評分和人工智能。每一個年代都有自己看來最客觀、最科學的工具。問題從來不是我們應否測量；沒有測量，很多判斷只能依靠印象和關係，未必更加公平。真正應該警惕的是：我們會不會忘記所有測量都只是從某一個角度觀看一個人？一個人的能力遠比一個數字複雜；一個人的將來，更不是一個模型可以完全預測。

三十五年前，一位專家曾經告訴我，我「不適宜讀書」。幸好，我沒有把那句話當成科學定律。今天假如同樣的判斷不是由一位專家說出來，而是由一個經過數百萬筆資料訓練的人工智能模型算出來，我們是否反而會更加不敢懷疑？想到這裏，我重新看文章的題目：「賢愚千載知誰是？」三十五年前，我以為這是在質疑那些替人判斷賢愚的專家；今天我覺得，它其實還可以多問一句：當判斷賢愚的權力由人交給機器之後，我們會不會因為機器看來更加客觀，反而忘記了繼續問——它憑甚麼這樣判斷？

[更多資訊](#)