米勒金字塔

前言

談到臨床能力的評估，一定會提到 George E. Miller 醫師刊載於 *Academic Medicine* 1990 年 9 月號的附刊（supplement）中的一篇文章"The Assessment of Clinical Skills/Competence/Performance"，這是他應邀年美國醫學院學會（Association of American Medical Colleges，AAMC）在舊金山舉行第 29 界年會演講的講稿。他在講稿中提出一個三角圖形的臨床評估框架（framework of clinical assessment），並稱它為金字塔，此後大家便稱此三角圖形為「米勒金字塔」（Miller Pyramid）。雖然許多人都引用這個圖形，甚至在圖上「塗鴉」成為新的版本，但真的有細心閱讀 Miller 的文章的人可能只是少數，大多似是看圖猜義，引發許多誤導，這恐怕是不負責任的行為。筆者特將原稿翻譯（如文後附件），並撰寫本文將 Miller 的理念作較精細及明確的介紹。

Miller 對金字塔的詮釋

Miller 在講稿中提到，這個演講主辦單位原意是請他介紹標準化病人（standardized patients，SP）在評估學生臨床能力相關研究的狀況，Miller 在介紹相關主題之前用了約五分之一的篇幅提出他個人對於臨床能力評估的架構，並繪了一個三角圖形來幫助說明，這原屬講題的開場白，後來這段概說竟在日後被醫學教育界奉為圭臬，恐怕連 Miller 本人也始料未及。

Miller 在他的講稿中對這「金字塔」有很明確的詮釋（圖 1）。他將臨床能力評估分為四個層級，從底層往上依序是：知道（knows）、知道如何（knows how）、展示如何（shows how）、作為（does）。四個層級的意涵分述如下：

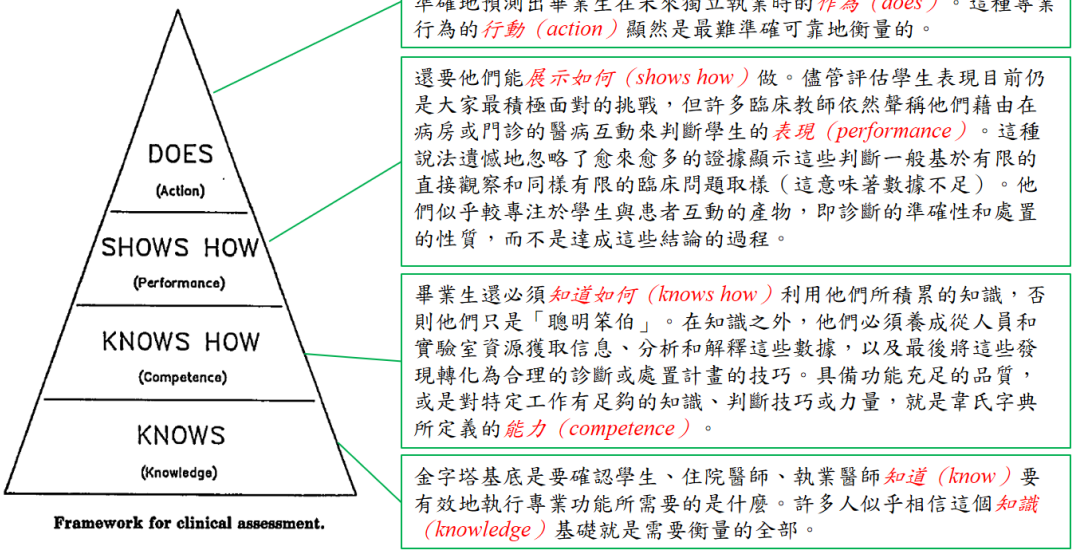

圖 1. 圖左為 George Miller 醫師所繪米勒金字塔的原圖。圖右為他在講稿中對每一評估層級的描述。

1. **Knows**：測驗學生是否具備有效地執行醫療專業所需要的知識（knowledge）。透過客觀的測驗方法（主要是選擇題）評量知識是目前考試機構和專科學會所採用考試系統的主流。對知識的測驗當然重要，但醫療作業不是光憑知識，所以單純的知識評估是不完整的考試工具。
2. **Knows How**：測驗學生是否知道如何應用他們的知識，包括蒐集、分析和解釋訊息和數據，以及將這些發現轉化為合理的診斷或處置計畫。對特定工作有足夠的知識、判斷技巧或力量，Miller 認為就是能力（competence）。
3. **Shows How**：測驗學生是否能展示如何去做。Miller 認為這個層級就是學生的表現（performance），是 Miller 這個講題的重點。在 1990 年，相關評估方法正在積極發展中（Miller 指的是利用 SP 來作評估以及當時發展仍未十分成熟的客觀結構式臨床測驗〔objective structured clinical examination，OSCE〕）。然而，當時仍有許多臨床教師不了解 SP 或 OSCE，依然聲稱他們藉由在病房或門診的醫病互動來判斷學生的表現。Miller 認為在臨床情境的直接觀察作為 shows how 的評估有三個問題有待解決：(1) 評分毫無標準，過於主觀；(2) 專注於學生與病人互動的產物（即診斷的準確性和處置的品質），而不是達成這些結論的過程；(3) 取樣不具代表性。其後由美國內科醫學會 Norcini 等人於 1995 年發表在 Ann Intern Med 有關簡短式臨床評量演練（mini-clinical evaluation exercise，mini-CEX）的研究報告，正是要克服這三個問題—有固定評分表、評分項目以跟病人互動過程為主、在不同場合作多次評估。因此，mini-CEX 和後續的直接觀察程序操作技巧（direct observation of procedural skills，DOPS）等工作場所導向評估（workplace-based assessment，WPBA）都是繼 OSCE 之後發展出來用作評估臨床表現的方法。
4. **Does**：測驗學生在未來獨立執業時的作為（does）。這是指專業行為的行動（action），Miller 認為無法在人為安排的考試場所中進行，是最難準確可靠地衡量的層級。Miller 可能並不知道早在二次大戰期間德軍即採用多源回饋評量軍官的表現，1950 年代埃索工程研究公司（Esso Research and Engineering Company）率先使用 360 度評估；他也可能不了解在二十世紀三十年代開始萌芽，五、六十年代才較為成形的學習歷程檔案（portfolio）。這並非 Miller 知這得太少，而是 360 度評估和學習歷程檔案這些評估方法在二十世紀九十年代之後才逐漸成熟。

擅作修訂之種種謬誤

在 2011 年 4 月 24 日中時電子報刊載「假病人出招 準醫師緊張」一文，記述國內醫學院首度聯合試辦醫學系應屆畢業生 OSCE 聯考首日狀況，經過一番讚揚這項創舉之後，出人意表地記述一位民眾蘇先生的看法：想要透過術科考試篩選出兼具醫術及醫德的好醫師，不僅是多此一舉，更是緣木求魚，因為明知是考試，再怎麼不耐煩也會「演」出愛心啊！有這樣的陳述並不奇怪，因為受訪者和採訪（報導）者都不了解 OSCE 是評估“shows how”。但蘇先生卻比許多做教育的人高明，如果考生明知是考試，再怎麼弄也無法評估“does”層面，無論 mini-CEX

或 DOPS 這些 WPBA，都是考生明知是考試，當然評的絕不是“does”層面。而 Miller 本人也認為在臨床上直接觀察是“shows how”，如果沒有處理好他所擔心的三個問題，WPBA 甚至不是好的評估方法！

古今早外，以訛傳訛是常見的事，輕則無傷大雅，重則誤導眾生，遺害久遠。然而，在醫學教界有關米勒金字塔的訊息，以訛傳訛的程度令人感到訝異。在網路上隨處可見相互引用的錯誤訊息（圖 2）。特別要聲明，這裡所謂錯誤的定義是指打著 Miller Pyramid 的旗號卻違背 Miller 原意的種種內容。

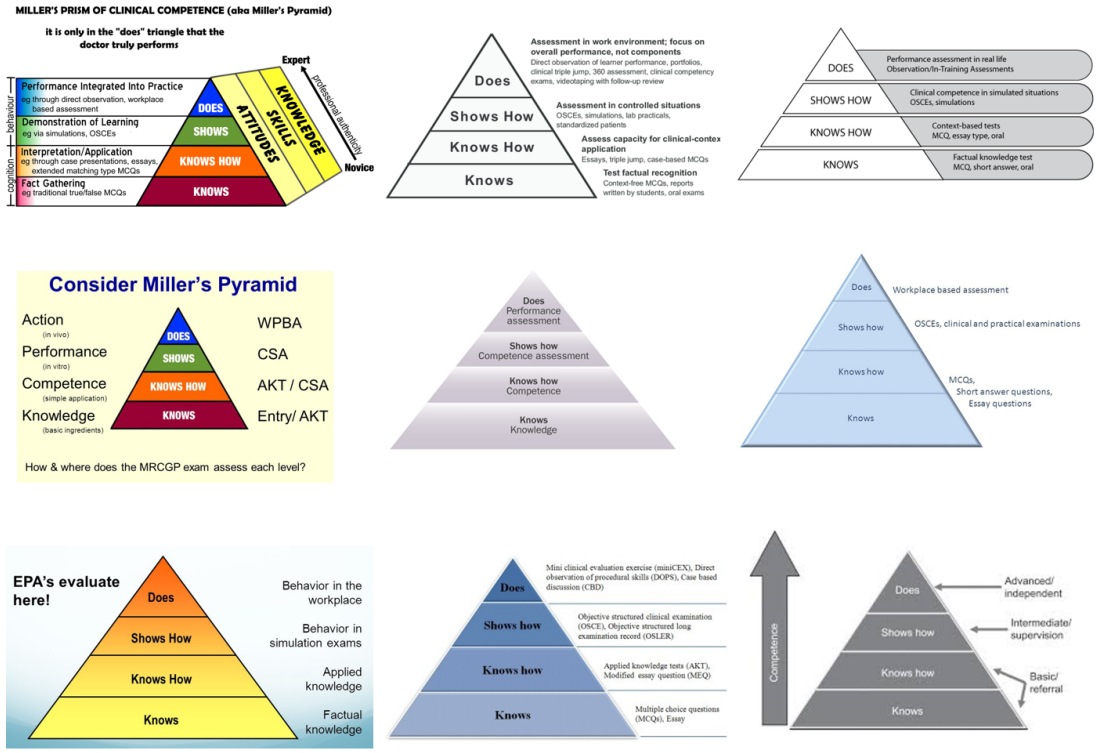

圖 2. 在網路上有關米勒金字塔的錯誤訊息俯拾即是，圖中九個例子為隨機取得，因為都有錯誤，不可仿效或參考，故不引用出處。

在眾多未遵照 Miller 原意而錯誤得最突兀的是在圖 2 左上角顏色頗為鮮豔的圖形，由於引用的人還不在少數，故在此特別指出其偏差，希望大家不要被誤導。修改者除了別出心裁地把 Miller 叫金字塔的三角形擅自改為棱形（prism），對原創者不甚尊重之外，新增的錯誤還有下列十項（圖 3）：

1. 將 Does 和 Shows how 層級同列測驗行為（behaviour），這不但與 Miller 原意相違（應為 Does 層級的 Action 才是行為），連民眾蘇先生都知道屬於 Shows how 的 OSCE 不能測驗學生的行為，這個錯誤有些不可思議。
2. 將 Does 視為“performance integrated into practice”不符 Miller 所認為的“action”。出現這謬誤的理由就是誤以為 WPBA 屬於“Does”層級。
3. 誤把 WPBA 放到“Does”層級。

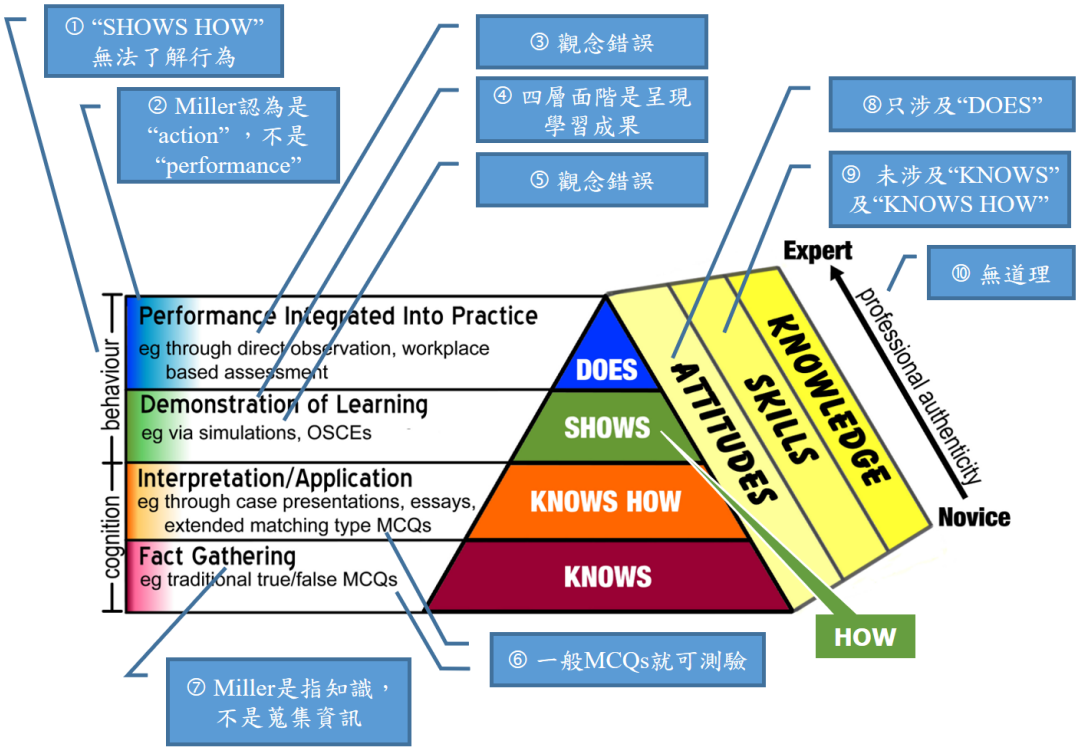

圖 3. 目前已知錯誤最多的米勒金字塔修改版本。

4. 把"Demonstration of Learning"列為"Shows how"專屬，事實上四個層級都涉及"Demonstration of Learning"。
5. 誤以為"Shows how"只能在模擬環境中進行。
6. 傳統是非題考"Knows"而延伸配對題考"Knows how"毫無道理，一般選擇題既可考"Knows"也可考"Knows how"。
7. 誤將"Knows"侷限為蒐集資訊，Miller 指的就是知識。
8. 誤將態度跨越四個層級，其實只有"Does"測驗態度。
9. 誤將技巧跨越四個層級，應未涉及"Knows"和"Knows how"。
10. 誤解四個層級是測驗不同階段和能力的考生，更誤導愈上層愈顯出專業特質，事實上四個層級涉及的是不同的測驗領域而非不同的程度。

Miller Pyramid 的解讀

Miller 的四個層級中，最上層的 Does 涉及知識、技能和態度，是不在特別安排的情況下評估，既考能與不能，也知為與不為。測驗方式包括學習歷程檔案、病歷紀錄抽查、學習護照及 360 度評估等（圖 4，A & B）。下一層 Shows how 則是只考能與不能，不知為與不為。測驗方式包括 OSCE、Mini-CEX、DOPS 等（圖 4，A & C）。金字塔基層的兩項（Knows 及 Knows how）可整合為一，是只考知與不知，推測能與不能。測驗方式包括病例報告、申論題、口試、Case-based discussion、簡答題、填充題、選擇題、是非題等（圖 4，A & D）。

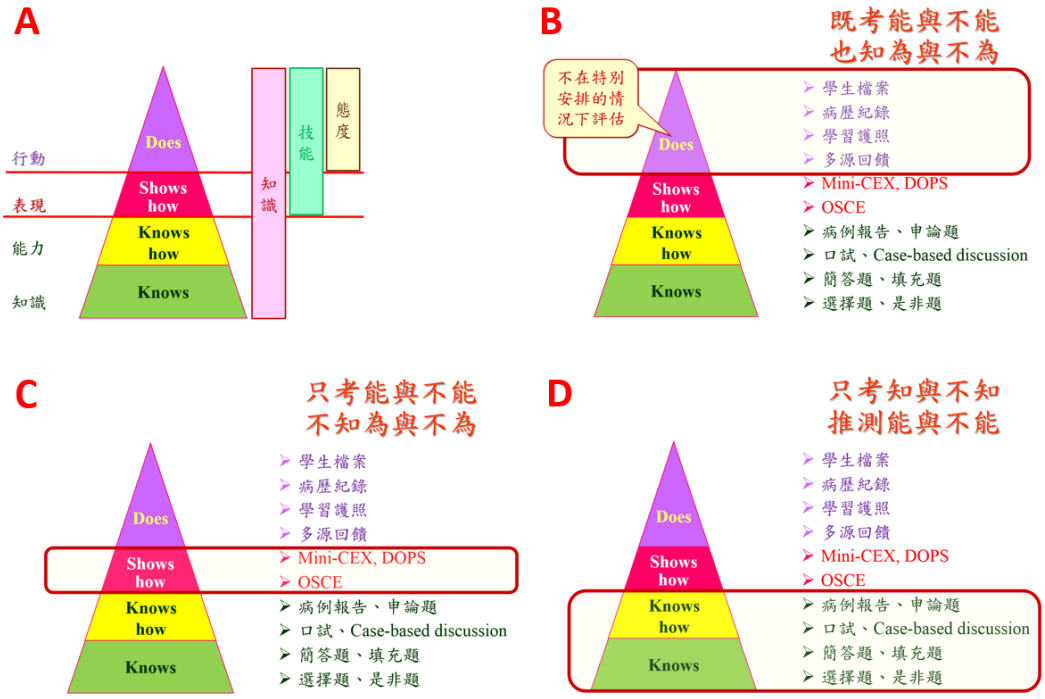

圖 4. Miller Pyramid 的解讀。

結語

George Miller 醫師投身醫學教育研究數十年，建立了伊利諾伊大學芝加哥分校醫學院的醫學教育部，在他退休之年的一個演講，似是無心插柳，讓他繪出的三角形對現今的醫學教育界產生深遠的影響。然而，卻因不少後人的不用心，將他原創的理念畫蛇添足，發生誤導，實在令人遺憾！Miller 的事件提醒我們務必提高警覺，對於新接觸的理念不可盲從，要以辨思甚至是批判的態度來學習，才不至失去方向。

附件：Dr. George E. Miller 講稿 “The Assessment of Clinical Skills/Competence/ Performance” 全文中譯版。

臨床技能/能力/表現評估

George E. Miller, M.D.

Acad Med 1990; 65: S63-S67.

剛好在 20 年前第八屆醫學教育研究（RIME）年會，我應邀演講當時稱為「醫學教育研究的視角」的講題。現在，已有十多年沒有從事第一線的教育研究，這個演講對我是一個高度的肯定，但也為我帶來一些不安，因為在場有這麼多比我更有資格和經驗的人在聆聽：例如 David Swanson、Geoff Norman、Paula Stillman 或者 Howard Barrows。然而，主辦者已經做出了他們的選擇，原因不太清楚，可能是要請一個年逾古稀的人再一次提供觀點。至少這是我會儘量去做的。

儘管有人建議將介紹重點放在標準化病人，但重要的是要從正確的認知開始，這就是：評估一位醫師的各項專業服務這樣複雜的事，沒有單一的評估方法可以提供判斷所需的所有數據。所以我首先建議一個評估需要用到的框架。

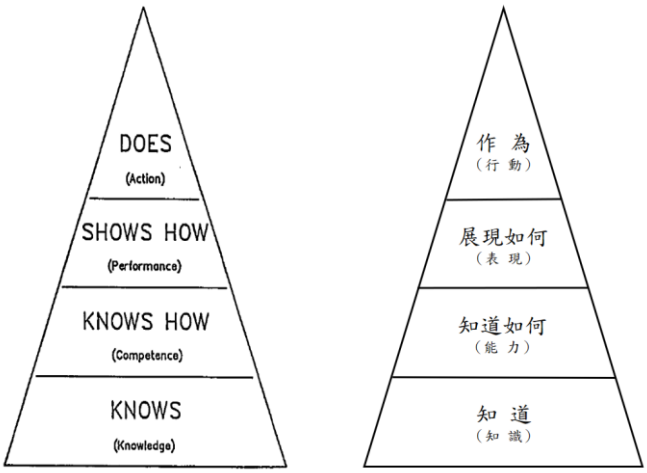

Figure 1. Framework for clinical assessment.　*圖 1.* 臨床評估框架

我使用這金字塔是要幫助說明（圖 1），它的基底是要確認學生、住院醫師、執業醫師*知道*（know）要有效地執行專業功能所需要的是什麼。許多人似乎相信這個*知識*（knowledge）基礎就是所需要衡量的全部。毫無疑問，評量知識（主要是透過客觀的測驗方法）是目前機構和專科學會考試系統的主流。

但正如許多年前 Alfred North Whitehead 指出，沒有什麼比僅僅知情的人更無用的了。對知識的測驗當然重要，但如果我們真的相信醫療作業不是光憑知識，就了解這些測驗是不完整的工具。

為了實現這個更廣泛的目標，畢業生還必須*知道如何*（know how）利用他們所積累的知識，否則他們只是個「聰明笨伯」。在知識之外，他們必須養成從人員和實驗室資源獲取信息、分析和解釋這些數據，以及最後將這些發現轉化為合理的診斷或處置計畫的技巧。具備功能充足的品質，或是對特定工作有足夠的知識、判斷技巧或力量，就是韋氏字典所定義的*能力*（competence）。

儘管檢測這些特質的測驗程序有重大進展，但是懷疑論者仍然指出，這樣的學術考試不能

證明學生在面對病人時會做什麼，即不僅要證明他們*知道（know）*和*知道如何（knowhow）*做，還要他們能*展示如何（shows how）*做。儘管評估學生表現目前仍是大家最積極面對的挑戰，但許多臨床教師依然聲稱他們藉由在病房或門診的醫病互動來判斷學生的*表現（performance）*。很遺憾，這種說法忽略了愈來愈多的證據顯示這些判斷一般基於有限的直接觀察和有限的臨床問題取樣（這意味著數據不足）。他們似乎較專注於學生與病人互動的產物，即診斷的準確性和處置的品質，而不是達成這些結論的過程。

最後，仍然令人存疑的是在一般人為的考試設置中所進行的評估，是否能夠準確地預測畢業生在未來獨立執業時的*作為（does）*。這種專業行為的*行動（action）*顯然是最難準確可靠地衡量的。近年來為完善評估系統的最後階段所付出的辛勤努力產生了不一致的結果，人們還必須繼續保持不減的活力。

與此同時，雖然可以合理地假設行動和表現是有賴這三角形基礎層面所達成，但測量這基礎部分（即知識和能力）並不能完全有把握地預測其上層更複雜的目標。當這個事實與考試推動教育制度不可避免地相結合時，因為它們以最清晰和最現實的語言傳達給學生如要成功必須學習或做哪些事情，那麼隨之而來的是教師應該同時尋求這個三角形的上層的教學方法和評估程序。

考慮到這種多維複雜性，讓我們來看看對於學生、住院醫師或執業醫師我們所知道的評定專業行為的個別評估技術。首先，對知識的評估，特別是通過客觀的測驗方法早有深入的研究，結果亦為眾所周知，這裡只需要簡要的陳述就可以了。這些已用得熟練的方法有很高的信度和抽樣有效性，而在實際的使用方面亦證明了它們的效用，儘管它們評估的範疇可能是有限的。

下一個層面是評估應用知識的智能技巧或執行診斷和治療程序的操作技能，開始出現某種程度的不確定性。採用教育目標的 Bloom 氏分類法作為準備選擇題試題的指引，確實有助於提升命題技巧，測驗的不僅是要回憶信息的片段。然而，仍然存在一些意見分歧，例如，如果一項旨在評估分析、解釋或合成的考題，在不知道考生以前是否經歷過類似的挑戰的情況下，是否可以自信地評定考生能達成這些項目？對於初學者來說原來可能需要一些較高層級過程的考驗，若考生已有過這樣的經歷，可能只需要簡單的回憶。

為了因應這種反對意見，考試經常使用以系列格式呈現的修訂版主題問答題（modified essay questions，MEQ）或病人處置問題（patient management problems，PMP）。這些考題每一題都先介紹一段臨床狀況為後續行動作開場。在下一步要解決問題的之前，在前段提供確認的或修改後的數據資料，而後段則沒有提供這樣的回饋，隨後的步驟是取決於初始的介入的效果。相對標準化的 MEQ 普遍有相當高的信度，而 PMP 的評分則存在問題。其中包括個別評分者之間就正面或負面權重分配每一個可能的介入措施，甚至是應該遵循的最佳路徑，難以達成共識。在準備充分時，有明確無誤的計分鍵和訓練有素的計分員，在一個考試中採用相當數量的 MEQ 和 PMP，信度大致相當。然而，對於 PMP 還有一個干擾狀況，就是紙本考題是不可避免地暴露出提示。

對這類考題的種種問題人們已努力解決，並透過更先進的技術來擴大它的使用範圍。其中最突出的是國家醫師考試委員會的電腦測驗計畫，該計畫包括臨床模擬和選擇題。新發展的電腦化模擬可隨著時間以動態方式引出考生須要分析和處理臨床問題的狀況。這些考試資料庫正在穩步地擴展，目前有超過 70 所醫學院正在嚴格審查中。

儘管存在著心理方面的問題，這些程序具有較高的表面效度，即與較簡單的技術相比更接

近真實的表現和行動要求，所以在某些情況下，稍為犧牲信度是可被接受的。然而，這些實例仍然是激烈辯論的問題，目前仍持續致力改善這些高效率機械式可計分測試形式的信度。

較無爭議的是應用模型在評估執行特定程序的技術能力的角色。雖然這些設施用於教學多於評估，但作為評估工具，它們在學生所面臨的挑戰中具有穩定性和一致性的優點。常用的包括急救安妮、生殖器及直腸和乳房模型、檢查眼底或耳膜的模型、簡單的心音模擬器，或更複雜的心血管系統模擬器—Harvey。這些品項的缺點與其設計的精準度不太相關，問題在於評分所需的檢核表和評分標準以及使用這些評分表的評分員。儘管不能完全克服，這些障礙可藉由仔細設計評分工具和訓練評分員而大為減少。

然而，這些方法之中每一種都欠缺與人互動的步驟。正因如此，許多教師堅持在與學生或住院醫師一起工作過程中，在病房、門診和其他臨床教學地點進行評估。不可否認的是，這相較於人為和獨立場境評估學生表現的任何設施更接近獨立執業的真實情境。對於這種評量程序的支持者普遍未能承認的是缺乏標準化、取樣的局限性，以及不常對表現本身的觀察 (不若對結果的討論) 以提供作出判斷的基本數據。它本質上是一種依靠臨床印象而不是系統性積累可靠信息的方法。

由受過訓練的評估者使用標準化的檢核表或評定量表直接觀察病史和身體診察確實解決了信度問題，但卻處理不了採樣問題，這是要獲得關於能力的一般結論至關重要的。偶爾應用它在造就性評量可能有很大的用處，但是它對總結性評估則明顯受限。

使用替代病人可對一些與真實臨床互動相關的複雜心理問題得到答案，　讓我們進入金字塔結構的下一個步驟。在這方面首先作出努力之一者是美國急診醫學會推出的角色扮演。

在這裡，醫師考官將被規範要描述特定病人問題的病史特徵，並根據要求提供關於身體和實驗室檢查結果的準確信息。在扮演的角色中，他們可以進一步對考生進行口試，然後用預先確定的和標準化的標準對整體表現進行判斷。雖然有證據顯示這些技術提供的觀點是傳統的方法不能獲得的，但同樣顯而易見的是，這種大規模的考試在金錢和人力上都是昂貴的。

對於特定的技術操作程序，另一種方法是聘用非醫師的婦科和泌尿科教學助理，對其進行生殖器和直腸檢查，並且能對操作技巧的準確性以及檢查者對病人舒適和理解的敏感度提供即時回饋。雖然最常用於教學目的，但這些人也已經成功地接受培訓，使用檢核表或評分量表來評估考生的表現。

但最有效替代真實情境的可能是使用標準化病人（SP）的模擬臨床互動。 當 Howard Barrows 在二十多年前引入這種正常的、訓練有素的模擬者時，他們準確地模仿異常臨床狀態的能力普遍遭受懷疑。我當時是懷疑者之一，但在我第一次觀摩的幾分鐘之內就抹去我的懷疑。目前你們大多數人可能都有類似的經驗，並有類似的反應。當然，大量的學生、住院醫師和執業醫師在回想起來表示，考間內的模擬病人與真實在醫院門診或私人診所的病人無法區分。

現在很清楚，對於誰可以受訓為模擬病人幾乎沒有限制，至少對於模擬病史、情緒狀態、種族和文化差異或病人類型的溝通部分是如此。模擬情境可以直接面對面、以電話交換意見，也可透過第三者來處理嬰幼兒、無意識病人或家庭的問題。

即使是一些不尋常的身體異常也可以透過最有天賦的標準化病人成功地模擬，如反射異常、抽搐、異常步態、發熱疼痛的關節，以及胸廓擴張受限等。但對於那些無法模擬的事物，許多研究者使用了具有穩定身體異常的真實病人，訓練他們提供符合這些發現的標準化病史。

但正如與單一病人的互動不能直接用來對整體臨床表現作一般性結論一樣，與標準化病人的單次互動也不能達到這個目的。適當的取樣議題仍然必須處理。 大約十年前，蘇格蘭丹地大學的 Ronald Harden 導入客觀結構式臨床測驗（OSCE），作為在合理時間內增加臨床行為樣本的一種方法，使用的設施和資源在大多數醫學院校都有提供。

Harden 採用的模式就是大家熟悉在解剖學和病理學實驗考試所用的多站測驗。在這個臨床版本的考站中，例如可包括對病人進行重點病史詢問或身體診察（由一個或多個考官評分）; 判讀 X 光或顯微鏡圖片或心電圖（並以書面作答）; 臨床數據分析，以及作出診斷或治療的結論（並評估其對書面問題的回答）。由於多站模式已經被許多其他團隊進一步利用，真正的病人常常被標準化病人所取代，以保證對考生的挑戰是一致的。所有這些都意味著 OSCE 本身並不是一種考試技術，而是一種可以採用各種技術（從選擇題到模擬）的格式。

醫學教育者擔心以循傳統方式了解知識的獲取作為臨床表現的評定，其感受的壓力愈來愈大，因而導致愈來愈多的醫學院校在其教學計畫中採用標準化病人或替代病人的方法。一項在 1988 年醫學教育評鑑委員會（LCME）的問卷調查顯示，97 所美國學校現在使用婦科或泌尿科教學助理，61 所使用標準化病人進行其他臨床技能指導。雖然在該調查中沒有記錄，但從其他來源的合理推斷，該方法大部分使用在臨床醫學入門課程。有 41 所學校也採用此方法評估臨床技能，其中超過半數使用它們來決定晉升或畢業。預計來年有愈來愈的課程將會使用。

當這些程序是用於教學，心理議題可能只是一個小問題；當被用於造就性評估也只會有輕微不安，但當標準化病人被用作總結性評估工具時，他們便是非常重要。當這些模擬方法用於認證或執照考試時便有較高風，問題將進一步加劇。那麼，在這個相對較早的發展階段，這些議題能談什麼呢？在此我將以很大程度地引用 Karl Ven Vleuten 和 David Swanson 付梓中的一篇非常精彩的評論。

當引入任何評估技術時，首先提出的問題之一就是測量的信度。在這種多站模式中，由於缺乏評分者間的一致性、標準化病人表現的落差，或者各個站之間的考生表現的差異，可能會影響使用標準化病人所獲得的分數的可重現性。對這些變數中的每一項都進行了或多或少的調查，但已經得出的結論仍然被認為是暫定的而有待進一步確認。

最初人們普遍認為，對考生表現做出公平判斷的觀察員必須是醫師，為了保證公平性和一致性，通常雇用兩名觀察員。如果這方法被廣泛使用，這種人力密集的方式會使可行性出現嚴重的問題。現在看來很清楚，當評估員有接受使用標準檢核表或評定量表的培訓，評估員間的一致性落在 0.5 到 0.9 之間，一般在 0.75 到 0.85 之間。這樣的情況在一般的考試時間長度，一名評估員會與兩名評估員沒有差別。在站數更多的時候才需要用到第二個評估員。

此外，已經發現，標準化病人本身或其他非醫師人員在使用精心設計的檢核表或評定量表進行適當的培訓時，可以像醫師那樣準確地描述考生的表現。醫學院教師是否會普遍接受這一發現並採取行動還有待觀察。

現在有愈來愈多的證據顯示，在單一地點接受培訓的幾名標準化病人演出相同角色時可達成可再現性。初步證據建議，在多個地點或不同訓練師進行培訓時，也可以實現這種一致性。雖然這可能不是個別機構關心的問題，但是當考慮到合作性的多機構測驗時，在經濟考量下這是具有很大的意義和重要性。

有一件事是不需要進一步的辯論：考生在一個案例的表現難以預測在其他案例的表現。內

容特異性的問題在這裡似乎和其他測驗方法一樣嚴重。現在看來，為了獲得可接受的可再現的分數，將需要至少三到四個小時的測驗時間。若使用 SP 的考站與數據解釋、鑑別診斷或實驗室技巧等一起測驗或隨後作答時，將需要更長的總測驗時間。這表示 SP 測驗最好用於評定直接與病人互動行為，而其他方面的臨床表現則採用更經濟的測驗方法進行評估。這樣劃分表現成分是否會扭曲專業行為的整體評估，需要進一步的研究。

關於最佳考站格式和強度已經有相當多的討論。從現在的證據可以得出結論，這應該由評量什麼事情來決定，而不是事先任意決定。較長的考站會提供更多的信息，但較短的考站將在同一時間段落中提供更廣泛的病人問題取樣。

最後應該指出的是，大多數信度的研究集中於分數的重現性而不是決策的重現性。目前還沒有大量工作用於製定 SP 測驗的絕對標準，但有一個強有力的論點是，將考生排名不是臨床表現評估的目標。真正的目標是確定是否達到了規定的掌握水平。如果將這種合格與不合格的觀點作為信度研究的重點，那麼可以預料，為了達到可支持的廣義結論，需要更短的測驗時間。這樣的焦點轉移也可提供機會來探索連串測驗的有用性，因為當大多數考生表現良好，那麼簡短的測驗將會可靠地認證大多數人，而詳細的評估則保留給那些表現品質有問題的人。

同樣重要的是效度問題。在此可以對這種品質的主觀評估充滿信心地說話，但對其經驗判斷的信心卻不足。當住院醫師和執業醫師在一系列臨床互動過程中遇到標準化病人而無法發現哪些人是真實的，哪些人是模擬的時，當然標準化病人必須具有高水平的表面效度(Geoff Normal 稱之為「信仰效度」)。而他們似乎也具有內容效度，因為考驗考生的表現是在醫學實踐中所要求的。這些行為的抽樣是否足夠或多樣，取決於設計藍圖的謹慎程度以及測驗與藍圖的匹配程度。但任何測驗都是如此。

迄今為止，以經驗證實的研究相對較少。那些已經進行的研究顯示有進階培訓的人表現優於初參與者，這些發現可以證實結構的有效性。而評定並行效度的研究亦進行中：較傳統的測驗的低相關性通常被作為不同品質被測量的證據，而由教師評分臨床表現的較高相關性則被作為兩者測量相同關鍵品質的證據。但是在每一個例子中都有另外兩個問題。首先是現在普遍接受的事實，即表現與知識不可分割，隨著教育的發展階段，知識可以增加，從而影響表現。其次，相關性研究通常是基於常模參照測驗的得分或排名，而不是標準參照評估的具體行為成就。

當某些操作元素得到特別處理時，標準化病人對於測驗的特殊貢獻就更為明顯。例如身體診察中的假陽性結果 (如心雜音、視神經乳頭水腫或關節積液)、或在沒有進行適當檢查卻提報出結果，這些情況雖不常見，但與可接受標準有顯著偏差，這些偏差卻可能會不被發現。這些不妥之處經常可被這些測驗技術所發現，就是發生在臨床輪換結束已經被教師鑑定為合格的學生。

採用 SP 的考試恆被提及的問題就是它的可行性。這是一個不能迴避的議題，但是由於沒有通用的費用評定方法，所以只能得出初步的結論。這些變數包括為數不少的 SP 培訓和使用成本 (以提供足以評估表現的抽樣數量)、開發案例和腳本以及檢核表和評定量表的時間和成本、考試所需的材料和供應成本、統計分數和報告結果的成本，以及判斷表現的醫師或非醫師 (即標準化病人或其他人) 的費用。若扣除開發的成本，目前每個學生參加完整的認證考試的費用估計為 100 到 200 美元之間。

然而，這樣的估計並不包括可能由幾所學校或測試機構的合作發展測試的潛在經濟模式。南伊利諾伊大學和麻薩諸塞大學都已經開始往這方面努力，並將在國家醫師考試委員會的鼓勵和

支持下執行進一步的合作研究。只有在目標是針對臨床表現進行總結性評估時，這些合作方案才可能符合經濟效益，儘管創建一個帶有相應腳本和檢核表或評量表的認證 SP 人才庫，最終也可能被證明是一個受歡迎的教學和造就性評估的資源。

因為 SP 測驗方法那樣的有前途和吸引力，預期這項技術將普遍應用於高風險的升級考試和認證程序，故必須對一些仍待解答的關鍵性問題作進一步研究。因為這是在座各位可能要承擔的工作，那麼讓我列舉一些似乎特別需要的研究。

最困難的問題之一是對採用 SP 的考試可用於評量專業行為的哪些部分達成共識。從現在使用的各種測試格式來看，很明顯的不同的教育團隊有不同的思維，這些差異可能對蒐集足夠的數據以獲得廣義結論所需的時間有顯著的影響。從成本效益的角度出發，如果資格考試採用 SP 的部分僅限於評估信息蒐集和溝通技巧（正如幾位著名的研究者所建議的那樣），或者是放棄一些重要元素，改由較傳統的測試方法評估及評定專業行為的其他層面，是否可行？

同樣令人困惑的問題是，什麼是標準化病人互動考題評分的最佳方法。這不僅是檢核表或評分量表，抑或由醫師或受過訓練的非醫師評分的問題，而是就觀察的相關層面達成一致看法，以及如何組合和加權這些觀察使能有意義地反映觀察到的表現的充分性。在回顧了許多目前正在使用的評分表格之後，van der Vleuten 和 Seanson 作出下列評論：「忽略重要項目並納入不重要項目的可能性非常高。前者懲罰作出未在評分表上的合宜行為的考生; 後者獎勵演出不合宜的考生。」

無論測驗的行為面向或所採用的評分程序為何，仍有一個尚未解決的問題，就是如何以最有效的方法制定表現的標準。過去，這個問題成功地以常模參照測驗來規避；但判定臨床表現似乎必須採用標準參照方法。在其他考試也曾在標準方面難以達成共識，而且沒有理由認為在使用 SP 的程序方面有任何不同。但是，如果我們要忠實於社會交付我們的責任，認證臨床表現是否足夠，而不僅僅是對考生排出名次，那麼我們不能再逃避責任，要去尋找一種方法來讓我們可以做到。

如果要成功地進行機構間合作，那麼至少還有一個問題須要解決：創建標準化病人庫以及技術和後勤的共享。麻薩諸塞大學和南伊利諾伊大學與其他地區性機構合作展開了一些關於這些問題的初步工作。前者培養了一批標準化病人，送往其他地點進行考試；後者制定了一套模擬案例和標準培訓程序，與另一所醫學院共享，以便他們能夠進行共同的考試。對於有限的實驗目標，這些程序中的每一項似乎都運作得很好。但是，如果要有更廣泛的共享，就必須找到令人信服的答案來解決一些仍然存在的問題。

例如，現在必須有比現時更有說服力的證據證明，由不同訓練師在同一地點或不同地點培訓的幾位標準化病人對特定病例的演出會有相當接近的表現。而且，如果同一 SP 重複演出，這位 SP 的表現是否經歷一段時間仍然穩定？沒有可比性和穩定性的評定紀錄，這考試程序的信度將受到嚴重的質疑。Robyn Tamblyn 的研究最近有令人鼓舞的發展，這不僅表明這個目標能夠實現，而且也提供了可能進一步提高可比性的方法。

雖然也許不那麼緊迫，但實踐效果仍然重要。與其他大多數測驗方法一樣，考生若有先前經驗通常能夠表現得更好。對於標準化病人來說應該也是如此，儘管事實上他們只是為了準確地呈現學生在醫院和門診中經常遇到的實際情況。

致力於實施採用 SP 認證考試的國家醫師考試委員會和外國醫學畢業生教育委員會，以及

可能開展這種努力的其他認證機構，特別關心的議題是成本和後勤事務。以目前的成本估算尚不清楚是否可以合理且精確地預測國內或國際區域進行大規模運作時的可能需求，因為任何計畫的發展階段耗費總會較多。即使估算是精確的，這樣的花費作為認證或考試費用合理嗎？如果有減少開支的方法，不管是什麼，也是值得開發的。在這一點上，最有希望的可能是循序考試策略，使用粗略篩選來識別所有明顯可以接受的考生，精細篩選則留作落在模糊的灰色地帶的考生。這項技術的研究是非常需要的，因為它對測驗新策略的最終實施具有重要的意義。

到目前為止，我試圖沈著下來，合理地表述對臨床技巧/能力/表現評估的認識狀況。但讓我以一套個人觀點來結束，有些人可能認為這些觀點不過是偏見。

首先是我感覺到這任務的緊迫性。長久以來，我們一直根據個別人員能否展現出具備相關團體 (最常見的是專科學會) 認為執業必需的知識來判斷他們是否準備好從事專業的執業。質疑知識的重要性是毫無意義的，儘管它的暫時性特質。更重要的是，我們果斷地以考試程序來證明知識還不足以成功地通過考試或作為醫生執行。能力測驗的每一次改良都是為了更接近目標，直到最近，採用 SP 考試的研究指出，我們致力模擬醫師與病人及家屬互動的情境已有一些成果，包括現實存在的所有含糊之處。

依目前的經驗和資料，這些方法似乎不僅是可取的，在解決目前的問題之後，應盡快推廣應用於一般或特殊執業的資格考試。

最後，在這些評估之中，是時候放棄常模參照程序的舒適偽裝，並採用標準參照測試。以任意截止點將考生排序，反映區別遠遠大於差異，既不是好的教育，也不是好的醫學。

說服保守的醫學院教師並不容易，以目前的作法，有系統地累積行為證據　會使印象逐漸改觀，讓這種變化循序漸進。雖然不能單靠數據說服他們，但沒有數據，如同一位敏銳的觀察家多年前指出的那樣，數據少而意見多，充滿熱情的論據肯定也會動搖。而這似乎描述了全球許多地方對臨床技巧/能力/表現評估的狀況。我只能希望，今天在場的醫學教育研究改革者，在這個問題上最終會得到 Adlai Stevenson 形容 Eleanor Roosevelt 的話 :「她寧願點燃一個詛咒黑暗的蠟燭，她的光芒已經溫暖了世界。」祝福您們在這個有價值的志業中有美好的成果。