教育學報，2022 年，第 50 卷第 2 期，頁 27–50


# 評估方式轉變對教師評分的影響及對策反思

廖 梁*
香港中文大學大學通識教育部

自 2015 年始，在香港的高等院校，學生學業評估由常模參照評估轉為標準參照評估。評估方式的轉變需要教師打破以往的評分習慣，對他們有一定挑戰。在標準參照評估實施初期，教師評分有兩個特點：一是評分依然受到常模評估方式影響，二是評分自主權有所提升。教師是否保留常模參照評分取決於院系能否就評估標準的使用和闡釋提供足夠的評分交流，以及院系對評估結果所採取的問責方式；教師對評分自主的要求則是標準參照評估本身的精神和教師對學術自由的捍衛所共同決定。要幫助教師更好地過渡至一種新的評估方式，需要關注學校問責、院系支持、教師學術自由三個層面並作反思。

關鍵詞：標準參照評估；問責；評分過程；評分自主權；學術自由

## 引 言

自 2015 年，應香港質素保證局要求，香港的高等院校就學生學業評估陸續採取標準參照評估（criterion-referenced assessment），此前各院校均採取常模參照評估（norm-referenced assessment）的方式。常模參照評估是根據學生在班級表現的優異程度進行成績評定，實際操作上學校會限制各等級人數，從而令整體的成績分布符合正態分布（Lok et al., 2016; Popham, 1978），因此這種評估方式又被稱為根據分數分布曲線而評分（grading on a bell curve）（Quality Assurance Council, 2015）。常模參照評估的理念在於鑑別和選拔人才，但自 20 世紀 60、70 年代，西方一些教育學家，例如 Glaser（1963）、Carroll（1963）、Bloom（1968）陸續對這種評估方式提出質疑，質疑聚焦於評估目的究竟在於區分學業表現高低還是了解學生掌握了哪些知識和技能？Bloom 和 Carroll 主張評估目的不應只是比較學業表現的優劣，還需了解具體的

* 通訊作者：廖梁（liaoliang@cuhk.edu.hk）

學習情況。Glaser 則首先提出標準參照評估的概念，[1] 指出要了解學生的學習程度，應採取一種新的學業評估方式；它通過制定一系列學生所能達到的知識和能力指標，確定一門具體科目的評估標準，教師不再根據分數分布，而是依據評估標準的表現評分和評定等級。標準參照評估目的在於通過評估標準（assessment criteria）的清晰化和公開化，使學生清楚自身的能力表現，進而建立起對學習過程的元認知（meta-cognition）和對學習進程的自我調節（self-regulating）（Sadler, 1989）。此外，相對於常模參照評估，標準參照評估更關注學生的「真實表現」，避免因為比較和競爭而令學生失去獲得「真實分數」的機會，以促進學生學習的主動性，加強學生之間的合作和互助。應該說，它是一種比常模參照更為強調「以學生為中心」（student center）和「以評促學」（assessment for learning）的評估理念。

然而在實踐時，由常模參照評估轉向標準參照評估面臨諸多挑戰。對許多教師來說，常模參照評估的「風險」遠小於標準參照評估。採取常模參照時，每個等級的人數會有相應規定，這樣出來的分數雖然不一定反映學生的真實表現，但至少表面看來符合分數分布要求，不至於出現大量高分或大量不合格的「反常」現象。假如要完全轉向標準參照評估，相當於失去了「等級限制」的安全欄。完全依照評估標準評分並非易事：一方面，常模參照評估的慣性難以去除，以香港高等院校實施標準參照的經驗來看，[2] 在相當長的時間內，無論是校方問責還是教師評分仍然容易受常模評分思維影響；另一方面，許多教師質疑，標準參照評估是否會造成學業評估的過分主觀化，從而導致高分泛濫（grade inflation）？

雖然標準參照評估解除了等級約束，賦予了教師更大的評分權，但教師所需擔負的責任亦更大。教師如何依據評估標準的要求給出公平的評分，如何吸納標準參照的內核理念從而令評估真正實現「以評促學」，無論對於教師、學生、學校或學業評估品質的影響都是深遠的。要令教師更好地使用標準參照評估，首先需要了解他們在評估轉變期所面臨的問題並思考相應對策。本研究探討了在常模參照評估轉向標準參照評估初期，教師最關注哪些問題？又有哪些因素會影響教師應用評估標準？這些因素如何影響評分過程？借助透視這些問題，文章討論了從哪些層面為教師所面臨的改變作準備，以幫助教師和學校更好地適應和運用這種新的評估方式。

## 文獻回顧

### 評分過程的複雜性

評分的重要性不言而喻，但關於教師在評分過程中的心理決策的相關研究則較少（Joughin et al., 2017）。部分學者認為評分是教師的專業判斷，具個人化特徵（Grainger et al., 2008），屬默會知識的一部分，它是隱晦且不被闡釋的（Sadler, 2005）。教師

在評分過程中使用的默會知識又稱為內在標準框架（internal framework）（Bloxham, den-Outer, et al., 2016）或隱性標準（tacit criteria）（Liao, 2022; Sadler, 2005），它在評分過程所起的作用甚至比外部評估標準更大。在 Bloxham, Boyd, et al.（2011）的實證研究中，她們採取放聲思考的方法記錄教師的評分過程，發現不同教師存在不同的內在標準框架，包括對評估標準的解釋不同、只使用部分評估標準、使用另外的評估標準代替現有標準等。研究者進一步指出，內在標準框架決定了教師對評估標準的闡釋、理解和應用，令評分行為呈現個體差異。

另一些學者則從社會情境脈絡觀察教師的評分行為，認為評分不是孤立的個人行動，而是受環境因素的影響（Rust et al., 2005; Shay, 2004, 2005）。南非學者 Shay（2004）結合布迪厄（Pierre Bourdieu）的社會實踐理論，研究了教師評分過程中的影響因素。她指出教師評分是社會文化脈絡下的情境式闡釋行為（socially situated interpretive act），任何跟評估活動相關的持份者和環境均會對評分產生影響，評估行為產生於這些人和物構成的「場域」（field）和「習慣」（habitus）之中。教師的評分不僅與個人判斷有關，還與學校願景、院系文化、同行關係、學生印象這些因素密切相關。

無論評分屬於個人化行為還是社會化行為，由於受到默會知識和社會情境脈絡的雙重影響，評分過程通常都隱晦而難以理解。評估過程的不透明會加大對評估的問責難度（Bloxham & Boyd, 2012; Sadler, 2017），校方與教師之間就評分結果的處理亦時有矛盾（Leathwood, 2005）。Hill（2011）總結了幾起關於成績評定的司法糾紛案件，發現其中的矛盾均源自校方和教師對評分權力的爭議。校方對教師評分權干涉過多會引起教師不滿，但如果學校對評分結果採取無為態度，又會使得評估質素受到影響。評估過程的不透明亦會加深學生對評估看法的隔閡。諸多研究顯示學生對評估結果的滿意度不高（McMorran et al., 2017），認為評估標準不明確（賈周聖，2014），師生之間缺乏針對評估的溝通（Brown & Wang, 2016），成績結果不是由個人能力或努力決定，而是由教師個人意願決定（O'Hagan & Wigglesworth, 2015）等等。

不少學者試圖從不同角度解釋造成評估過程晦澀、缺乏透明的原因。有學者指出教師通常基於自身經驗、專業訓練、教育價值理念等理解評估標準，從而造成闡釋和評分的差異（Bloxham, den-Outer, et al., 2016; Sadler, 2009a）；亦有學者則從評估的判斷方式上對教師評分的差異加以解釋。Sadler（2005, 2009b）指出教師在使用評分量表（grading rubric）時有兩種取向：一是根據每條評估標準的要求評分，以得出最終分數，這種方式叫分析式判斷（analytic judgement）；另一則是根據學生表現給出一個綜合評分，這種方式叫整體式判斷（holistic judgement）。整體式判斷與鑑賞藝術品的行為模式相似，教師通常基於「直覺」作判斷，並認為將評估分解為不同指標會破壞學生能力表現的整體性和連貫性。

對於標準參照評估而言，採取分析式判斷似乎理所當然，但實際情況卻不盡然。實證研究顯示，許多教師在基於量表評分時，並不會逐條按評估標準分析進而相加得出總成績，而是使用整體式判斷，基於「印象」和「直覺」給分（Grainger et al., 2008）。還有一些教師則是首先使用整體式判斷，再根據評分量表分配各評估指標上的分數（Bloxham, Boyd, et al., 2011）。這些研究結果正符合 Sadler（2009b）所描述的「雙重判斷方式」，即教師在成績判定中，實際上存在分析式和整體式的混合判斷，且教師並未清楚覺察自己究竟採取哪種判斷方式。無論是雙重判斷方式的存在，抑或判斷本身難以覺察的特點，都令原本就複雜且隱性的評分過程變得更加難以理解。

## 標準參照評估實施過程及問題

標準參照評估的核心是使用外在、確定的評估標準評定成績，因此其實施過程便是對評估標準的實踐應用。具體過程包括：確定評估標準、設計評分量表、使用評分量表、評分後針對評估標準做校準（moderation or calibration）（廖梁等，2021；Carlson et al., 2000）。社會建構視角[3]是標準參照評估實施中最常見的視角。例如 Rust et al.（2005）提出一個建構式的評估過程模型（social constructivist assessment process model），主張評估是教師和學生對評估標準的雙向建構過程。教師需要就評估標準的含義不斷討論和詮釋，並在教師之間形成實踐共同體（community of practice）（Bearman & Ajjawi, 2021; Wenger, 1998）；評估標準的內涵經過建構之後，再與學生就評估過程繼續溝通和解釋；最後根據評分結果的反饋，實踐共同體成員對評估標準進一步反思和修正。社會建構模式的意義在於注重教師對評估標準的再闡釋和修正，而不是將評估標準視為政策性、權威的、不可改變的事物。教師之間增強對評估標準的互動，可以減少對評分標準的個人化闡釋，使得評分更加公平；亦能改變評估標準在評分實踐中常被忽略的狀況，令外在評估標準發揮其應有的作用。

社會建構視角對教師如何詮釋評分標準，尤其是如何協調不同的個人化闡釋提出了思考。諸多研究者指出要協調個人化闡釋，需要針對外在評估標準建構「共同理解」（shared understanding）（O'Connell et al., 2016; Sadler, 2013; Watty et al., 2014）。「共同理解」能夠在評估標準的規範下減少教師之間的評分差異，從而提升評分信度（grade reliability）（Baume et al., 2004; Liao, 2022）。建構「共同理解」則是以探討評估過程為中心，在實踐共同體中為教師提供充足的評分交流機會；具體可以包括評分之前對評估標準的學習、認識、評分培訓等，以及評分之後通過校準對評估標準的使用進行討論（廖梁等，2021；Zahra et al., 2017）。然而，無論是評分之前的學習和交流，還是評分之後的校準和討論，在實踐中均遇到一定困難。這些困難既包括外在支持方面的缺乏，例如評分交流的機會不多、評分培訓的設計瑕疵、校準受到成績結果的影響、不夠聚焦於評分標準內容本身等，亦包括教師個人化闡釋致使難以

形成有效「共同理解」的問題（廖梁、梁美儀，2022；Liao, 2022; O'Connell et al., 2016; Zahra et al., 2017）。

## 教師在評估轉變中遇到的困難

要令教師改變已經習慣的評估方式，轉而使用新評估方式是評估改變（assessment change）中的一個難點。諸多因素揭示出評估轉變過程中的困難。例如，如果教師感覺被排除在評估實踐之外（Deneen & Boud, 2014），或者發現自身在評估中所能夠發揮的作用被壓縮（Simper, 2020），均會對新評估產生牴觸心理。此外，政策制定一方缺乏清晰指令，或者與教師溝通不佳，亦會影響教師對新評估的恰當使用（Brower et al., 2017）。拋開政策實施等外部影響因素，教師自身的因素亦加大了評估轉變的難度。有研究總結出五個阻礙評估改變的心理暗示，分別是：維持現狀心理、時間規劃、習慣影響、稟賦效應和宜家效應、對評分過於自信（Joughin et al., 2017）。這些心理歸根結柢在於教師並沒有真正了解新評估的要求，且對已經習慣的評估方式過於樂觀。在看不到潛在問題的情況下，教師對於改變自然缺乏動力。Joughin et al. 進一步指出，僅僅依靠教師單方面的力量並不能令評估行為發生改變，「要令評估行為發生改變，只考慮教師這一層面的做法非常普遍，但實際上這是錯誤的」（p. 1229），如何組織和實施評估理念，學校、院系在評估轉變中同樣發揮重要作用；如果只是採取政策傳達或者「例行公事」般監督的做法，只會令教師抱怨，「除非將評估置於教學中心地位，以及評估品質受高度重視，否則，自上而下的政策號召難以令教師改變評分習慣」（p. 1229）。

# 實證研究設計

## 研究目的和問題

對於所有香港高等院校而言，從常模參照評估轉向標準參照評估是學生學業成績評定中的共同經歷。從評分本身的複雜性、評估改變的難度和評估標準在實施中的問題來看，向標準參照評估的轉變並非水到渠成之事。本研究目的之一在於呈現評估轉變過程中教師所遇到的問題。此外，標準參照評估雖然在香港屬於較新的評估方式，但在西方一些高等院校，已經實施了較長時間。儘管標準參照評估的實施歷史並不短暫，但在實際評分時如何「恰當」使用評估標準，從已有文獻來看，還存在諸多需要審慎思考的問題，例如：如何看待外部評估標準與教師隱性標準之間的張力？教師的評分決策有多大程度受外部評估標準的影響？本研究目的之二在於通過探究教師與評估標準的互動，了解評分背後的考慮和決策，討論評分的影響因素。從實踐

層面而言，理解在轉變期所面臨的問題和評分的影響因素，能夠幫助教師、院系和大學更好地實施這種新的評估方式。從評估理論層面而言，這些問題的呈現和分析能揭示評估標準在評分運用中的複雜性，從而進一步探究標準參照評估有效實施的條件。具體而言，研究將回答以下問題：

1. 在向標準參照評估轉變的過程中，教師對標準參照評估的總體評價如何？感知到哪些實施中的困難？
2. 教師如何使用評估標準來評分？評分背後的考慮有哪些？
3. 評估轉變期中的評分呈現哪些特點？這些特點如何形成？
4. 如何幫助教師度過評估轉型期，並促進標準參照評估的實施？

## 研究背景

香港中文大學自 2018 年開始，全部院系和教學單位均由常模參照評估轉為標準參照評估。中文大學鼓勵各院系根據學科自身特徵，自行探索和確立適合課程要求的標準參照評估。研究資料基於對中文大學某一教學單位（下稱 A 部門）的觀察，以及對部分教師的訪談。A 部門為中文大學本科學生提供系列必修課程，選擇 A 部門為研究案例，主要基於三點：

1. 該部門所開設的必修課程是由二十多名全職教師組成的團隊共同完成，相對於專業科系一門課程通常只由一至兩名教師開設，這些必修課程涉及教師人數眾多，學科背景各異，能更廣泛地了解不同學科背景教師對評分方式轉變的觀感、應對和評價；
2. 在標準參照評估要求下，不同教師需要面對相同的評估標準和評分量表，這能更好地了解實踐共同體的建立以及在評分過程中所遇到的問題。不同教師對相同評估標準的不同闡釋和使用方式，亦能更好地揭示評分的差異；
3. 該部門的教師團隊經過多年磨合，在評分方面已經形成一定的氛圍，這種氛圍正是 Shay（2004）所描述的基於組織、文化、價值理念等各種不同場域所形成的某些慣性想法和做法，當評估方式轉變而需要打破之前的評分習慣時，在群體中觀察這種改變更能豐富地捕捉到不同層面（例如學校、院系、教師個人）對評分改變所能起到的作用。

儘管在此前的一些部門會議中，A 部門對標準參照評估曾有一定的討論，但真正落實這一政策是在 2018 年暑假。為應對秋季新學期向新評估方式的轉變，A 部門於暑假期間成立了工作小組（task force），由教師、課程主任和研究人員共六人組成，以研發新的評分量表。新評分量表中的評估標準選擇基於預期學習目標（intended learning outcomes），所體現的是學習成果取向（outcome-based approach），這一理念

是香港質素保證局對香港各高等院校的「課程與教學品質檢視」中標準參照評估的理念原型。在此之前，部門教師雖然有各自的評分量表，但它更多是基於寫作過程中的規範，而非基於預期學習目標。新的評分量表在評分標準和內容描述上均作出一定改變。此外，在評分量表的使用上，教師是否必須在評分時使用評分量表，在舊量表時期部門沒有強制規定，針對評分量表的公開討論亦很少。由於主要依照常模參照評分，教師在當時有各自的評分方案，例如百分制、五分制等等。新評分量表中的總分、各等級分數區間（score range）則是固定的，這需要教師改變原有的分數方案而使用相同的方案。

## 資料收集

資料收集分三個方面：

1. **觀察** —— 研究者參與了標準參照評估量表的開發（2018 年暑假）、關於評分量表使用的全體教師會議（2018 年暑假），以及評分後的校準（2019 年年中）。在整個過程中，研究者以參與式觀察者的身分，關注教師對新評估方式的反應和反饋，特別是教師關於評分背後的考慮和決策，以了解新評估自落實到實施、評估工作小結整個過程中，與教師的互動及其內涵。
2. **訪談** —— 為更深入了解教師的評分考慮，研究選擇 A 部門七位教師進行訪談。訪談時間在 2018 年年末，七位教師分別來自物理、生物、工程、中文、文化研究、歷史專業。選擇不同專業的教師是為了避免特定學科和專業對評分的影響。已有文獻顯示不同教學經驗的教師，其評分考慮有諸多不同（Shay, 2005），因此教學經驗亦是選擇受訪對象的考慮因素之一。七位教師中有五位是該課程成立以來便任教的「元老」教師，兩位教師任職經驗不超過 2 年。每位教師進行一對一個人訪談，訪談時間為 1.5–2 小時，具體訪談問題見附件。
3. **其他途徑** —— 資料收集還包括 A 部門有關標準參照評估的會議資料。正式會議總共兩次，持續時長為 1.5–2 小時。在會議中，筆者作為觀察者記錄了教師的提問和教師之間的互動。此外，還有會議之後公開討論的電郵資料，A 部門就標準參照評估實施展開的全校性教師意見調查（開放式問題）結果回饋，以及與教師的非正式談話。

## 資料分析

訪談及其他資料（例如電郵、意見調查、非正式談話後的反思筆記）均轉化為文本形式進行內容分析（content analysis）。其中，訪談資料因問題集中和內容強度，是資料分析的主體。研究者對文本資料進行了兩輪編碼。第一輪編碼圍繞訪談問題，

呈現出「標準參照評估與教師」、「教師對標準參照評估的實際應用」這兩個主題。其中「標準參照評估與教師」將關注點放在教師如何評價標準參照評估、如何理解標準參照評估的內涵要求，以及教師的擔憂這些問題上。「教師對標準參照評估的實際應用」則包括了教師如何闡釋評估標準、如何評價和使用評分量表，以及採取何種判斷方式來評分（這是依據 Sadler 對判斷方式的劃分，見前文）。在第二輪編碼中，研究者基於兩大主題及內容，提取和整合出描述評分過程和影響評分的三個重要事件，這三個重要事件可以幫助讀者了解教師在評估方式轉變中所面對的問題、如何評分和評分背後的考慮；相應地，如何促進教師評估轉變的反思和對策建議亦基於這三個事件之上。這三個事件為：在評估轉變期間教師所面對的問題和教師對新評估的看法；對評分量表的使用；問責方和專業自主對教師評分的影響。研究結果將圍繞這三個事件展開。

## 研究結果

### 轉變期教師所面對的問題和教師對新評估方式的看法

當評估轉為標準參照方式之後，教師首先關注的是自己能否適應這種改變。適應意味着需要時間和緩衝空間，而從評分量表的開發到真正實施僅有幾個月時間，恰恰反映預留給評估改變的時間準備並不足夠。當教師感到準備不足時，改變的動力往往急劇降低，例如關於分數方案的改變。由於採取標準參照評估之前，教師有各自的評分方案，而採取標準參照評估之後，要求所有教師使用同樣的評分方案，且每個等級的分數區間須一致。部分教師不贊成使用相同的評分方案，有教師認為應該給教師多一些時間適應：

> 還是應該給時間讓教師自己摸索一下，看哪種評分方案更合適。

有教師則表示轉變可以按照循序漸進方式進行：

> 標準參照評估與之前的評估方式相比有很大不同，如果評分方案變化少一些，對教師而言過渡可以順利一些。

在對標準參照評估概念的理解上，並不是所有教師都清楚常模參照與標準參照的區別，這一點在全校教師問答調查中尤為明顯。有教師表示不清楚改用標準參照評分之後，是否還能自行調整分數分布。亦有教師不清楚標準參照的意義，認為「即使有評估標準，評分主要還是依靠教師的自我判斷」。有的教師則表示還是可以跟隨常模參照的做法，依據分數分布來評分，以簡化評分過程：

是否可以把事情簡單化一些，依然按照曲線上的分數分布來評分呢？

在深入的個人訪談中，教師並未像意見調查中顯示的那樣，對標準參照和常模參照做法模糊化。所有參與訪談的七名教師均能較準確地說出標準參照評估的特點，例如「不再以分數分布為評分依據，而是依據學生真實看法」；「能夠減少學生之間的互相比較，促進合作式學習」。值得一提的是，雖然受訪教師能夠較好地把握標準參照評估的內涵，大多數受訪者並未對是否繼續保留常模參照表達明確的態度。一些教師在闡釋和評價標準參照評估時是中立甚至正面的，但當進一步探究他們如何使用標準參照對學生作業評分時，發現依據分數曲線評分依然扮演了重要的調節作用。例如有教師會預估學生的能力分布，再根據這種心目中的分布來評分，這實際是一種更為隱性的分數曲線評分：

其實最後〔標準參照評估〕出來的分數結果和之前的分數分布相比不會相差太大，……通常能力比較突出的學生大致佔 17–20%，而大約 70% 的學生能力一般，剩下的那部分學生通常學習態度有一定問題。

這種隱性的分數曲線評分具體表現在教師並不是完全按照標準的描述評分，而是結合心目中的分數分布對評分作微調：

我會按照實際做一些調試，例如可能各項都不錯便可以是 A，而不是每項都達到量表的要求。

我會根據中期成績評定結果〔決定期末論文評分〕，如果中期成績和分數分布預設相差太大（例如 30% 在 A），我會調整期末論文的評分。

總體而言，時間準備是教師在面對評估方式轉變中的第一個要求，繼而是教師對標準參照的看法，這會直接影響教師的評分實踐。在轉變初期，舊的評分方式影響較突出，許多教師仍希望能夠沿用常模參照，或者採取將兩種評估結合的方式。這種對改變的牴觸有時候以表面順從的方式進行，即一方面使用評估標準評分，但同時卻依據分數曲線調整最終的成績結果。

## 使用評分量表及其背後的評分考慮

在使用評分量表時，教師首先關注的是評估標準的描述。參與訪談的教師總體對評估標準的描述持肯定態度，認為標準能夠幫助他們區分不同能力等級的學業表現。但亦有教師提到表徵學習能力的關鍵字匯比較抽象，例如有一條標準描述為「展示出

深思熟慮的個人見解以及有洞見的反思」，何謂「深思熟慮的個人見解」和「有洞見的反思」？教師認為僅憑這些描述並不能作出相應的判斷，還需要進一步的闡釋。有教師這樣說：

> 採取標準參照評估的話，就需要提供具體明確的評分標準。雖然評分量表提供了一些描述，但詞彙量太少了，例如展現出個人反思，甚麼是個人反思呢？我自己會進一步補充對評估標準的闡釋。

另外一個普遍的困惑是對各等級標準描述的「度」的把握。例如「優秀的理解能力」和「足夠的理解能力」中「優秀」和「足夠」如何區分的問題。七位教師中有四位提到評估標準的設定，尤其是對等級 A 的設定偏嚴格。有教師提及寫作評估標準更適合於學期末的論文評核，而對於學期中段的反思日記而言，則要求頗高：

> 學生的能力是逐漸形成的，在學期中段很難考核到一些高階思維能力，例如對主題的深度探索。我在反思日記中不會對學生有過多高階思維的要求，只需要能夠很好地解釋文本內容即可。所以在反思日記的考核中我需要〔對量表〕做一些調整。

七位教師並不全是採取分析式評分，有兩位表示會使用整體式評分方式，一位表示會使用混合式評分，即先用分析式評分，之後將自己的「印象分」與評分比較，再決定是否需要調整分數。教師對分析式評分的顧慮在於不確定評分結果的「準確度」，具體而言是指將每項評估指標的得分相加而得到的總分數是否和自己的原初判斷一致？這一看法背後反映出在常模評分中，教師可能更多是採取整體式評分，而轉為標準參照評分之後，評估指標的設立決定了評分需轉向分析式，這對不同教師會造成不同程度的適應問題。有教師表示自己還是更加適應整體式評分的方式：

> 我有嘗試採取分析式評分的方式，結果發現評分速度大大減慢。將各項指標相加有時會把我弄糊塗，最後得出的結果未必符合我的〔最初〕判斷。

在使用評估標準評分時，還有一個爭議是關於是否需要使用相同的分數區間。部分教師認為應使用一致的分數區間，即 A 等級的分數在 80–100 等等；另一部分教師則認為可以保留教師原先的分數區間。有教師認為標準參照評估只是要求採取相同的評估標準，而不是相同的分數方案，在評分方案上教師可以有自己的選擇：

> 我理解大家只需在評分中統一評分標準就好，但不需要使用一樣的分數區間。用甚麼評分方案還是應該尊重教師個人的評分權力。

總體而言，轉變為標準參照評估之後，評分量表的重要性相對常模評估時期得以突顯。教師使用評估標準時，會關注標準的內容描述是否具有區分度、要求是否合理、內容是否明確等等。但在標準參照評估初期，教師最關心的還是如何評分的問題。如何評分在教師當中存在一些爭議，是採取分析式評分還是整體式評分？是否需要保留教師個人的評分方案？這些爭議歸根結柢源於評估方式的轉變。失去了常模評分的安全欄，教師不確定評分結果的「合理性」，亦因此尤為關注如何評分，以及評分是否符合自己的「真實」判斷。

## 轉變期中的評分特點

評分自主性有所提升在教師使用評估標準時體現較為明顯，主要表現在兩方面。一是當教師對評估標準的內容有疑問或者不確定時，會根據自我判斷採取行動，例如調整標準的鬆緊程度，進一步闡釋標準的內容含義等。調整評分量表，使其更加符合自身對考核任務的設定，在教師之中十分普遍：

> ……平時的反思日記，我對他們的要求只是能夠很好地理解文本就可以了，達不到其他思維能力諸如多角度看待問題、展示問題複雜性這些，那可能學生做到「好」我已經認為是 A 了。

> 例如當他回答比較全面，可能就可以去到 B 或者 B+，但如果還能有一些洞見，則成績會更好。

二是教師傾向採取自己熟悉的評分方式和評分方案，例如採取整體式而非分析式評分，以及使用自己劃定的各等級分數區間。值得一提的是，有經驗的教師與新手教師在評分中所展示的自主決策有較大分別。新手教師對於自己的評分決策是否與評估標準吻合表現出更多的不確定；有新手教師認為在評分之前，應請評估專家為評分進行培訓：

> 一些 sub-grade 像 A–和 B+這些，對我而言還是比較難判定的。如果有一些培訓，或者邀請一些評估方面的專家給大家講一下評分當中的技巧就好了。這樣我覺得我的評分也會更準確一些。

有經驗的教師則對自己的評分決策更有自信。雖然有經驗的教師亦會提出諸如評估標準內容較抽象等問題，但對如何準確闡釋評估標準並沒有過多疑慮。調整評分鬆緊、策略性闡釋評估標準以滿足不同考核任務（例如中期考核和期末考核）的要求、採取整體式評分方式等，有經驗的教師能更自如、更自信地運用這些策略。

雖然教師相信個人的專業決策能力，但同時亦很關心行政上如何問責評分結果。這一點教師在個人訪談中雖然未有主動談及，但在教師意見調查和交流會議中，最引起關注的問題之一就是如何監管評估結果。有教師問到：

> 實行標準參照是不是意味着不會再看分數分布了？
>
> 今後如何保障分數公平呢？例如分數膨脹？

有教師一方面希望取消分數分布問責，從而令評分更寬鬆，另一方面又對學校真會這麼做表示懷疑：

> 為何學校評估委員會依然在使用舊的分數分布做指引？學校是否真的會放鬆對等級的限制？

有教師批評院系依然採取分數分布的做法違背了標準參照的原則：

> 為何〔院系領導〕還是要求我不能給太多 A？這不是違背了標準參照的原則嗎？

教師均關心分數分布是否徹底告別歷史舞台，繼續以分數分布為問責內容對教師評分影響較大。由於在標準參照評估實施上，中文大學鼓勵各院系和教學單位採用適合自身學科特色的方式，各院系在監督評分結果方面的做法是不一致的，而沿用分數分布問責則可能或多或少在各院系中均存在。在分數分布依然需要發揮作用的情況下，如何善用分數分布就構成了問題的關鍵。例如 A 部門對評分監管的做法是根據分數分布結果觀察評分中的異常情況（例如多 A 或者多不合格），再在出現分數異常的成績中抽選一些學生作業，結合教師提供的評分量表，請同專業的教師（匿名）重新根據評分量表評定學生成績，以確認評分是否公平。A 部門負責課程評估的領導認為：

> 教師只要評分有依據，多 A 是沒有問題的，只要教師對評分做出合理的解釋即可。但我們也會用匿名覆核的方式來確保學生是否真的受到公平對待。如果匿名覆核成績結果發現兩者評分相差很大，則教師除了需要解釋評分之外，還需要根據匿名覆核建議做一定的修改。

總體而言，轉為標準參照評估之後，教師對評估標準的使用雖然有堅持專業自主的一面，但依然受到分數分布的影響。

# 總結、討論與對策

## 研究結果總結

在向新評估方式轉變的過程中，教師所面對的問題包括：時間準備、對標準參照評估意義的把握、對評估標準內容的闡釋，以及如何綜合各條評估標準評分。其中，時間準備和意義把握存在相互影響。如果預留給教師了解和適應新評估的時間過於倉卒，則不利於教師充分了解標準參照評估所要實現的教育理念。如果教師不能準確區分標準參照評估和常模參照評估的內涵差異，便很難真正在評分中充分運用評估標準，評分亦更容易受舊模式和習慣影響。評分轉變期需要面對的問題還包括：對新評估標準的闡釋和使用評估標準進行評分。由於評估標準不是定量式的內容描述，教師需要進一步闡釋其內容含義。如何對標準中的一些術語作出準確闡釋，對教師（尤其是新手教師）而言有頗多不確定感。

總體而言，教師在轉變時期的評分呈現兩個特點。一是評分受到舊評分方式（即常模評分）影響較大。研究發現根據分數分布評分是評估轉變期最常見的一種隱性標準。常模參照對教師評分的影響十分複雜：一方面，當教師對新評估本身存在不確定性，例如無法獲知自身對評估標準的闡釋是否「適當」時，常常採取常模評分的方法以避免因為對評估標準闡釋不當而出現分數分布異常的現象；另一方面，即使教師有信心對評估標準作出合理闡釋，其評分決策並不一定完全依照評估標準，而是隱性地使用分數分布，研究推測這主要與對評分的問責內容有關。

二是評分自主性有所提高。評分自主性體現在教師會主動調試評分量表，對評估標準靈活闡釋等等。尤其在轉變初期，A 部門在組織教師評分交流方面的活動比較缺乏，教師只能依靠自身對評估標準的理解和評分經驗來評分，這就更加強化了教師在評分量表和評估標準闡釋方面的自我主導。研究亦發現，有經驗教師的評分自主性明顯高於新手教師，而新手教師則希望獲得更多評分培訓、評分交流等支持。另外，評分自主性還體現在教師不希望問責過多地受分數分布所制約。

## 討 論

新評估方式的轉變並不是一蹴而就的，評估轉變期會產生一系列問題。轉型期評分的第一個特點是教師評分受到舊的常模參照評估影響，部分教師將按照曲線評分視為「安全合理」的做法，亦有教師會潛在地使用分數分布評分以調節最終成績。研究認為，教師保留或者潛在使用分數分布，主要受兩方面的影響。其一是教師之間缺乏對如何使用評估標準的共同理解。由於對評估標準的理解和使用主要基於個人闡釋，這使得教師無法準確估計自我闡釋的「準確度」，從而引起對評分可能招致的負面後果的不確定性。在這種情況下，教師轉而尋求更安全的「分數分布」就變得

可以理解了。因此，學校和院系如何就標準參照評估的實施來敘事，將極大影響教師是否真正使用評估標準。如前文所述，實踐共同體是幫助教師建構「共同理解」的重要條件。實踐共同體的構建需要院系對轉變期中的評估準備作必要的指導、培訓和評分交流，尤其對教學經驗尚不豐富的新手教師而言，評分之前需獲得更多關於評估理論和實踐方面的支持。

另外一個保留或者潛在使用分數分布的影響來自問責內容。如何認定一個評估結果是「可接受」或者「好」，值得行政層面的深度審思。行政層面對成績評定的評估恰恰構成了前文所述評分過程中的「場域」，它對塑造評分「習慣」有不可低估的作用。從受訪教師和全校意見調查來看，不少院系仍以分數分布為主要問責內容。研究認為，如果院系只依照分數曲線檢視評估結果，會令教師看不到制定評估標準的意義何在，則很容易出現表面順從的情況，即雖然提供了評分量表，但教師不會嚴格按照評分量表評分，亦不會與學生解釋評分量表和評分依據。當某個等級出現人數超量的現象，教師會在遵照標準參照的精神和安全評分之間權衡，而最終選擇安全評分的教師應該不會是少數。

轉型期評分的第二個特點是評分自主權的提升。研究發現，雖然以分數分布為問責內容以及實踐共同體的缺失均會令教師被動地指向常模參照，但標準參照評估的實施亦同時促進了教師評分自主權的提升。教師最關注的是如何評分才能令評分結果「滿意」。滿意通常包括了自我、問責和學生三方。對教師而言，首先達到自我滿意居於三方中最重要的位置。通過觀察研究發現，教師在評分決策時，雖然會同時考慮行政問責和學生需求，但評分首先要令自己「信服」是大多數教師的主要考慮。例如教師對如何使用評分量表有着自己的堅持；面對分數分布的問責方式，有教師表達了個人評分不應該受到約束的看法。

獲得足夠的評分自主權，是令自我滿意的前提條件。評分自主權表現在教師認為自己有權利決定如何使用評估標準，例如把握評估標準的「度」以及對評分量表作調試，教師視這些內在隱性標準的使用為理所當然。相應地，極少有教師意識到就評分一起集體討論的意義。當這種自主意識過強時，有些教師甚至會視評估標準為一種擺設，認為評分歸根結柢還是教師的主觀判斷。評分自主權還表現在評分方式和評分方案的選擇上，教師傾向於選擇自己習慣的、而非符合標準參照精神的方式和方案，評分自主權還表現在最終呈現的成績結果上。以分數分布為問責內容的方式令部分教師對分數調整產生「慣性」，但另一部分教師則希望通過標準參照的實施令自己獲得更大的評分自由。

值得注意的是，雖然諸多研究指出，評分自主權彰顯了學術自由（academic freedom），增加了教師主動尋求改變評估的動力（Deneen & Boud, 2014; Simper, 2020），但研究亦發現，評分自主權導致了教師在標準理解、評分方式、鬆緊程度上

的把握均不一致。這些差異容易導致評分的差異化（O'Connell et al., 2016）以及成績結果的信度不高（Baird et al., 2004; Yorke, 2011）。評分差異和信度問題是大學生學業評估（尤其是針對開放性問題，例如寫作）的常見問題（Baume et al., 2004）。評分差異大、信度低則易引起前文所說的負面效應，評估信度低亦會影響校方與教師之間的關係。如前文 Hill（2011）所回顧的幾個司法案例中，教師認為自己有足夠的評分權，校方則認為自己有成績的最終決定權，而校方行使決定權的方式則使教師修改分數或者乾脆由校方代勞，引發了教師對學術自由的擔憂和不滿（Buglear, 2011; Sadler, 2011）。

歸納而言，院系在創建實踐共同體方面能提供的支持及資源、問責方式及內容、教師對評分自主權的看法，共同影響了教師將在成績評定中如何使用評估標準，亦影響了標準參照評估的有效實施。這些影響因素呼應了 Hill（2011）指出評估過程的「張力三角」。從教師角度而言，無論是遵從外部標準、外部評分規則還是對評估結果進行干預，都影響其評分自主。校方視防止評分過於主觀、保證評分公平、確保每位學生受到公平對待為基本責任。而評分的公平、盡量避免個人主觀評分造成的「偏見」問題，則於學生至關重要。評估過程缺乏透明度、基於評估標準而建立的學習策略和學習回饋缺失，除了如前文所述對學生有負面影響，亦可能導致學生對學習轉向功利主義，變得只是一味追求分數（林銀玲、葉信治，2014）。

## 對 策

從教師嘗試標準參照評估的經驗，可以看到評估政策在行動中的轉變不是自然而然便可發生，校方、院系、教師之間需要相互溝通和支援，僅僅依靠教師個人無法真正實現評估轉變。

首先，質素保證局和校方是評估政策的制定者、實施方和問責方。質素保證局需要對標準參照評估的目標、理念、實施提供更多解釋、說明和指導，只是採取自上而下的方式便會出現如研究中所顯示的各種問題，例如教師對評估標準解釋不一、使用不當、採取各自的評分方案等等。在問責方面，學校應該就如何審核成績結果與教師充分溝通，而不是繼續照搬常模參照的做法，只關注分數分布。這種只根據結果、模型、慣例、統計數字等等的問責方式只是如前文所回顧的技術理性式的做法，它通常會忽視評估過程中的各種問題，與標準參照精神不符。如文初指出，標準參照評估的理念在於幫助學生了解和掌握自身的學習能力和學習狀態。這意味着評估最重要的目的不在於出示一個無意義的分數或者等級符號，而是通過評估促進學生對自我能力的覺察，並形成付出努力達至目標的精神。要令這一理念得以彰顯，需要問責時以評估標準和評估過程為本，而不是只簡單追求「理想」的分數分布。

其次，院系屬政策的另一實施方，需要為教師提供足夠的評估資源和交流平台。學生學業評估向來是高等教育中一個薄弱的環節（Knight, 2002），相對於科研的專業化和教學的頻繁互動，學業評估幾乎處於邊緣地位。一方面，教師缺乏專業的評估知識，評分主要依靠自身經驗積累（Sadler, 2011）；另一方面，教師認為評估是自己的「個人事務」。受訪教師中所體現的對評估標準的個人闡釋、對評分量表的策略性調試等等，正是將評分「個人事務化」的體現。它的負面效應在於這些內隱化的個人標準將導致評分過程不透明，容易遮掩評分中的主觀性問題。院系需要為教師（尤其是新手教師）提供成績評定方面的理論和實踐經驗，還需要就評估組織多次集體討論。這些討論可以令教師看到彼此對評估標準的不同看法。置身於實踐共同體之中，教師能夠自我調整，從而構建對評估標準的「共同理解」（Sadler, 2013; Wenger, 1998; Zahra et al., 2017）。

最後，評估順利轉變的關鍵依然是評估的實踐者 —— 教師。除了在問責上擺脫技術理性的傾向，在評估過程中構建實踐共同體，還應該關注教師的學術自由。許多國家已把教師的學術自由保障寫入了教育法案，例如 1988 年英國的教育改革法案（Hill, 2011）和聯合國的「高等教育指南」（United Nations Educational, Scientific and Cultural Organization, 1998），美國則把教師的評分權視為言論自由的一部分，受憲法第一修正案保護（Fossey, 2007）。然而，理應受到保護的評分權和學術自由，並不總是產生好的效果，評分結果的公平性一直被詬病。有學者指出教師在評分中所施展的權力，正侵害了學生的利益。學生為了獲得好成績，往往在學業考核中盡力表現教師所希望的那一面，而忽略對問題的真正思考（Lee, 2006）。

問題的關鍵在於教師如何看待評分所體現的學術自由。標準參照擺脫了常模參照依照成績排名而評分的制約，表面看來似乎令教師獲得了更大的評分自由。但真正有意義的是教師如何合理應用這種評分自由，這將決定學業評估是否一場公正且為學生提供寶貴學習經驗的教學活動。假如教師將評分自主權等同於按照自己的意志評分，如意見調查中所申述的「評估標準作用不大，最終還是教師自己決定怎麼評分」，那麼評分主觀性所導致的評分可信性（grade integrity）（Sadler, 2009a）將令人擔憂；即使如受訪教師所顯示，教師主要依據個人理解和闡釋去使用評估標準，缺乏與同行的溝通交流，亦會令評估主觀性的問題比經過共同理解的評估更加嚴重。

Berlin（1969）把自由分為兩種：積極自由（positive liberty）和消極自由（negative liberty）。積極自由強調在實踐共同體中建立共同目標，共同體成員為了達成共同目標而放棄個人好惡，從而為團隊的發展創造方向；消極自由則強調建立私人領域，且不希望個人行為受周圍的人所影響和妨礙，這樣最終只會導致自我封閉。真正需要尊重和發揚的是積極的評分自由。這需要教師在評分中盡量避免受其他因素的負面影響，對評分作出全面審視和客觀公正的判斷。只有教師能主動意識到評分自由不是任意

評分，而是在了解評估目的和考慮到評估對學生的影響下教師對課程實施和教學結果採取的審視行動，才能在保障評分自由和自主權和評分公平之間劃定出合理的邊界。評分行為不僅是對教師自我教學的檢閱，還體現了對學生的深切責任。

## 意義和貢獻

研究揭示了向標準參照評估轉變的過程中，教師對評分的理解以及評分決策所呈現的特點。標準參照評估在香港高等院校實施時間尚短，相關研究並不多見。歐美雖然不乏對如何運用評估標準的討論和研究，但專門針對轉變期（尤其是從常模參照過渡到標準參照)的案例很少。本研究為標準參照評估的實施提供了實踐案例和建議，為理解標準參照評估下的評分決策提供了理論闡釋，相信研究發現可以豐富學生學業評估的理論和實踐。

在評分所受影響方面，已有文獻多從評分主觀性、評分信度、評分差異等出發，探討了通過建立實踐共同體以促進評估者之間的「共同理解」，從而減小評分的主觀性。但如何建立實踐共同體，從哪個層面建立共同體，已有文獻討論不多。從本實證研究來看，僅由教師群體通過自發方式並不會自然產生實踐共同體。因為這涉及到交流組織、會議統籌、選定討論主題等管理層面的安排，而教師一方面因為自身教學、科研、評估等事務，精力有限，另一方面在統籌同行共同活動上亦缺乏相應的行政支持。因此，院系層面為教師建立實踐共同體，提供評估所需要的學習資源，是較為恰當的做法。

問責內容和方式對教師評分有較大影響，這是相關研究領域裏的新發現。研究指出，當評分轉為標準參照評估之後，不適宜再繼續用分數分布評估成績評定的好壞。院系和學校需要採取與標準參照評估理念相匹配的問責方式，例如關注評估過程、關注教師和學生對評估標準的闡釋及理解等。

研究還揭示了轉為標準參照評估之後，教師對評分自主權的關注。雖然這並不是新的議題，但研究重新審視了如何看待評分權。關於教師在評分中所能體現的評分權力，或者說自由裁量權（discretion），已有研究多從教師權益角度肯定其重要意義（Deneen & Boud, 2014; Medland, 2016; Simper, 2020），但這樣的觀點往往伴隨對評估主觀性問題的「消極不作為」（廖梁、梁美儀，2022）。例如 Bloxham, den-Outer, et al.（2016）認為評估主觀性不可避免，「或許應該重新審視評估信度的問題……需誠實告之學生評估的複雜性，以及評估結果可能不準確」（p. 479）。這種看法雖然保護了教師的評分權和學術自由，卻迴避了教師可能因為個人偏見、與學生交往程度等因素造成的評分不公，實際上是抑制了學生所能獲得的學術自由。技術理性這種忽視教師個人決策的評估取向固然不可取，絕對的評分自由同樣不可取。在保護教師評分自由

的同時，教師自身亦需要有穩定持久的責任意識。這種責任意識在於實現評分的公平。教師需要警惕將評分變為一個人的孤島。教師可以與同行建構共同理解，這一過程亦可實現對自我評估標準的再理解。

## 註 釋

1. 標準參照評估最早由 Glaser 於 1963 年在《美國心理學家》（*American Psychologist*）期刊中提出。那時，評估主要採取測量（measurement）的形式，當時稱作標準參照測量（criterion-referenced measurement）。
2. 應香港質素保證局要求，自 2007 年始，香港的高等院校在學生學業評估上開始轉向標準參照評估，但直到 2015 年質素保證局第二輪課程與品質檢視後，各院校才真正開始實施標準參照評估。具體見質素保證局發布的「課程與教學品質檢視報告」（Quality Assurance Council, 2015）。
3. 與社會建構視角相對應的是技術理性視角（techno-rationalism），內涵主要包括：自上而下（top-to-down）的評估政策實施（Bloxham, Boyd, et al., 2011）；評估標準可客觀化（Medland, 2016）；評估標準的「語言描述」足夠構成對它的理解，毋須建構共同理解（Orr, 2007）。

## 參考文獻

林銀玲、葉信治（2014）。〈論表層學習與深層學習 —— 基於大學生學業評價制度改革的研究〉。《福建師範大學學報（哲學社會科學版）》，第 186 卷第 3 期，頁 151–156。

賈周聖（2014）。〈大學生學業評價改革研究〉。《教學研究》，第 37 卷第 2 期，頁 26–29。

廖梁、王永雄、彭金滿（2021）。〈標準參照評估的行動實踐 —— 以香港中文大學通識教育基礎課程為案例〉。《復旦教育論壇》，第 19 卷第 4 期，頁 52–59。

廖梁、梁美儀（2022）。〈標準參照評估的歷史發展及展望〉。《外國教育研究》，第 2 期，頁 60–76。

Baird, J., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 331–348. https://doi.org/10.1080/0969594042000304627

Baume, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment and Evaluation in Higher Education*, *29*(4), 451–477. https://doi.org/10.1080/02602930310001689037

Bearman, M., & Ajjawi, R. (2021). Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education*, *46*(2), 359–368. https://doi.org/10.1080/03075079.2019.1637842

Berlin, I. (1969). Two concepts of liberty. In I. Berlin, *Four essays on liberty* (pp. 118–172). Oxford University Press.

Bloom, B. S. (1968). *Learning for mastery*. University of California Press.

Bloxham, S., & Boyd, P. (2012). Accountability in grading student work: Securing academic standards in a twenty-first century quality assurance context. *British Educational Research Journal*, *38*(4), 615–634. https://doi.org/10.1080/01411926.2011.569007

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*(6), 655–670. https://doi.org/10.1080/03075071003777716

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment and Evaluation in Higher Education*, *41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Brower, R., Jones, T. B., Tandberg, D., Hu, S., & Park, T. (2017). Comprehensive developmental education reform in Florida: A policy implementation typology. *The Journal of Higher Education*, *88*(6), 809–834. https://doi.org/10.1080/00221546.2016.1272091

Brown, G. T. L., & Wang, Z. (2016). Understanding Chinese university student conceptions of assessment: Cultural similarities and jurisdictional differences between Hong Kong and China. *Social Psychology of Education*, *19*(1), 151–173. https://doi.org/10.1007/s11218-015-9322-x

Buglear, J. (2011). Grading and academic freedom: An English academic's angle on Hill's contentious triangle. *Quality in Higher Education*, *17*(1), 101–104. https://doi.org/10.1080/13538322.2011.554633

Carlson, T., MacDonald, D., Gorely, T., Hanrahan, S., & Burgess-Limerick, R. (2000). Implementing criterion-referenced assessment within a multi-disciplinary university department. *Higher Education Research and Development*, *19*(1), 103–116. https://doi.org/10.1080/07294360050020507

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*(8), 723–733. https://doi.org/10.1177/016146816306400801

Deneen, C., & Boud, D. (2014). Patterns of resistance in managing assessment change. *Assessment and Evaluation in Higher Education*, *39*(5), 577–591. https://doi.org/10.1080/02602938.2013.859654

Fossey, R. (2007). University oversight of professors' teaching activities: A professor's academic freedom does not mean freedom from institutional regulation. *Journal of Personnel Evaluation in Education*, *19*(3–4), 159–173. https://doi.org/10.1007/s11092-007-9043-6

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*(8), 519–521. https://doi.org/10.1037/h0049294

Grainger, P., Purnell, K., & Zipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment and Evaluation in Higher Education*, *33*(2), 133–142. https://doi.org/10.1080/02602930601125681

Hill, D. (2011). A contentious triangle: Grading and academic freedom in the academy. *Higher Education Quarterly*, *65*(1), 3–11. https://doi.org/10.1111/j.1468-2273.2010.00465.x

Joughin, G., Dawson, P., & Boud, D. (2017). Improving assessment tasks through addressing our unconscious limits to change. *Assessment and Evaluation in Higher Education*, *42*(8), 1221–1232. https://doi.org/10.1080/02602938.2016.1257689

Knight, P. T. (2002). The Achilles' heel of quality: The assessment of student learning. *Quality in Higher Education*, *8*(1), 107–115. https://doi.org/10.1080/13538320220127506

Leathwood, C. (2005). Assessment policy and practice in higher education: Purpose, standards and equity. *Assessment and Evaluation in Higher Education*, *30*(3), 307–324. https://doi.org/10.1080/02602930500063876

Lee, D. E. (2006). Academic freedom, critical thinking and teaching ethics. *Arts and Humanities in Higher Education*, *5*(2), 199–208. https://doi.org/10.1177/1474022206064037

Liao, L. (2022). Dancing with explicit criteria or marginalising them: The complexity of grading student work and the reconstruction of the meaning of criterion-referenced assessment. *Teaching in Higher Education*. https://doi.org/10.1080/13562517.2022.2119076

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment and Evaluation in Higher Education*, *41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

McMorran, C., Ragupathi, K., & Luo, S. (2017). Assessment and learning without grades? Motivations and concerns with implementing gradeless learning in higher education. *Assessment and Evaluation in Higher Education*, *42*(3), 361–377. https://doi.org/10.1080/02602938.2015.1114584

Medland, E. (2016). Assessment in higher education: Drivers, barriers and directions for change in the UK. *Assessment and Evaluation in Higher Education*, *41*(1), 81–96. https://doi.org/10.1080/02602938.2014.982072

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment and Evaluation in Higher Education*, *41*(3), 331–349. https://doi.org/10.1080/02602938.2015.1008398

O'Hagan, S. R., &Wigglesworth, G. (2015). Who's marking my essay? The assessment of non-native speaker and native-speaker undergraduate essays in an Australian higher education context. *Studies in Higher Education*, *40*(9), 1729–1747. https://doi.org/10.1080/03075079.2014.896890

Orr, S. (2007). Assessment moderation: Constructing the marks and constructing the students. *Assessment and Evaluation in Higher Education*, *32*(6), 645–656. https://doi.org/10.1080/02602930601117068

Popham, W. J. (1978). *Criterion-referenced measurement*. Prentice Hall.

Quality Assurance Council. (2015). *Report of a quality audit of The Chinese University of Hong Kong*. https://www.ugc.edu.hk/doc/eng/qac/report/cuhk201510e.pdf

Rust, C., O'Donovan, B., & Price, M. (2005). A social constructivist assessment process model: How the research literature shows us this could be best practice. *Assessment and Evaluation in Higher Education*, *30*(3), 231–240. https://doi.org/10.1080/02602930500063819

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, *30*(2), 175–194. https://doi.org/10.1080/0260293042000264262

Sadler, D. R. (2009a). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*(7), 807–826. https://doi.org/10.1080/03075070802706553

Sadler, D. R. (2009b). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education*, *34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Sadler, D. R. (2011). Academic freedom, achievement standards and professional identity. *Quality in Higher Education*, *17*(1), 85–100. https://doi.org/10.1080/13538322.2011.554639

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy and Practice*, *20*(1), 5–19. https://doi.org/10.1080/0969594X.2012.714742

Sadler, D. R. (2017). Academic achievement standards and quality assurance. *Quality in Higher Education*, *23*(2), 81–99. https://doi.org/10.1080/13538322.2017.1356614

Shay, S. (2004). The assessment of complex performance: A socially situated interpretive act. *Harvard Educational Review*, *74*(3), 307–329. https://doi.org/10.17763/haer.74.3.wq16l67103324520

Shay, S. (2005). The assessment of complex tasks: A double reading. *Studies in Higher Education*, *30*(6), 663–679. https://doi.org/10.1080/03075070500339988

Simper, N. (2020). Assessment thresholds for academic staff: Constructive alignment and differentiation of standards. *Assessment and Evaluation in Higher Education*, *45*(7), 1016–1030. https://doi.org/10.1080/02602938.2020.1718600

United Nations Educational, Scientific and Cultural Organization. (1998). *Records of the General Conference*. Author.

Watty, K., Freeman, M., Howieson, B., Hancock, P., O'Connell, P., De Lange, P., & Abraham, A. (2014). Social moderation, assessment and assuring standards for accounting graduates. *Assessment and Evaluation in Higher Education*, *39*(4), 461–478. https://doi.org/10.1080/02602938.2013.848336

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

Yorke, M. (2011). Summative assessment: Dealing with the “measurement fallacy.” *Studies in Higher Education*, *36*(3), 251–273. https://doi.org/10.1080/03075070903545082

Zahra, D., Robinson, I., Roberts, M., Coombes, L., Cockerill, J., & Burr, S. (2017). Rigour in moderation processes is more important than the choice of method. *Assessment and Evaluation in Higher Education*, *42*(7), 1159–1176. https://doi.org/10.1080/02602938.2016.1236183

# 附錄：開放式訪談提綱

## 關於評估改變

1. 你認為標準參照評估與之前的評估有何不同？
2. 你覺得標準參照評估的要點（核心概念）有哪些？
3. 採取標準參照評估對你而言，是否存在困難？如有，是甚麼？
4. 你是否會在課堂中與學生介紹這種新的評估方式，或者與他們一起討論評分量表？你為何這樣考慮？

## 關於評估標準的使用

1. 你是否認為評分量表具有良好的區分度（幫助你更好地判斷 A、B、C 不同等級）？
2. 你怎麼理解這些評估標準？它們（在評分上）好用嗎？
3. 你會如何使用評分量表去評定學生成績？比如你是根據逐條評估標準評分還是先給出一個總體印象分？
4. 你是否擔心使用標準參照評估後，會出現多 A 或者沒有 A 的現象？如果遇到這種情況，你會如何處理？

# Reflections on the Impacts of Grading Judgment Toward the Transition From Norm-referenced Assessment to Criterion-referenced Assessment

Liang LIAO

***Abstract***

*Since 2015, student assessment has been transferred from norm-referencing to criterion-referencing in all public universities of Hong Kong. During the transitional period, teachers have to face lots of challenges such as adapting to the new assessment and changing the previous grading habits. This study revealed two characteristics of grading behavior existing in the transitional period. On the one hand, grading has been greatly affected by the pattern of norm-referencing; on the other hand, demands for grading autonomy have increased comparing to the time when norm-referenced assessment was used. Whether teachers would go back to seek for the guide of norm-referencing is primarily depended on how the institution supervises the results of grades as well as whether adequate communications are created between department and teachers. Meanwhile, demands of grading autonomy are inspired by the spirit of criterion-referenced assessment as well as the desire for defending academic freedom from teachers. The study suggested that to facilitate the change of assessment and to better implement criterion-referenced assessment, supports, cooperation and reflections from different spheres of organizations and agencies (including the administration) are necessary, as well as its considerations on the grade accountability, resources and grading communication opportunities provided by departments, and the deep understanding of academic freedom from individual teachers.*

*Keywords: criterion-referenced assessment; accountability; grading process; grading autonomy; academic freedom*

**LIAO, Liang**（廖梁）is Research Associate in the Office of University General Education, The Chinese University of Hong Kong.