教育學報，2016 年，第 44 卷第 1 期，頁 195–212

《全港性系統評估檢討報告》的批判：政策評鑑研究的視域

曾榮光
香港中文大學教育行政與政策學系

2016 年 2 月 29 日，教育局在網上發布了《全港性系統評估檢討報告》（下稱《檢討報告》）。縱觀這份報告，都找不到任何政策檢討及評鑑研究應有的關於政策效果及／或後果的實徵分析及論證；然而報告卻能斷定可以「肯定系統評估的設立原意及價值」，並建議只要對實施細節稍作調整，就可以繼續「於 2017 年在全港實施」。本文正是要從公共政策檢討及評鑑的視域及方法出發，從三方面對《檢討報告》加以批判。第一，對《檢討報告》「肯定系統評估的設立原意及價值」是「回饋學與教」的立論加以質疑與批判。第二，從實徵角度論證實施了十年的系統評估根本就沒有實現《檢討報告》聲稱的「回饋學與教」的一種「進展性評估」應有的政策效果，即為未達標的學生提供「補底」的教學支援，相反卻對全港學校、教師、學生及家長造成巨大的應試操練壓力這種政策後果。最後，本文對《檢討報告》的主要政策建議更不表贊同，並認為根本沒有必要繼續每年強迫全港小三、小六及中三近十萬名學生參加這個評估。若單從系統性評估的角度，一個隨機抽樣的小樣本學校評估就可達到目的。

關鍵詞：全港性系統評估；進展性評估；總結性評估；質素保證機器；政策評鑑研究

2016 年 2 月 4 日，「基本能力評估及評估素養統籌委員會」的擴大工作組（下稱「工作組」）向教育局呈交了《全港性系統評估檢討報告》（下稱《檢討報告》）。這個由 35 位教育界精英組成的「工作組」，經過了 9 次會議，出席了 20 場持份者研討會，參考了 800 份意見書，「辛勤工作」而製作成報告書；[1] 教育局即時在官方網頁上同步發放了當中的「初步建議」，及後於 2 月 29 日更在網上發放報告全文，並開展《檢討報告》建議的跟進工作。至此，對教育局而言，應是對《檢討報告》毫無異議地照單全收，而對社會上種種對系統評估的批評亦算是作了回應，爭議似乎可告一段落。

然而，就嚴格的政策檢討和評鑑（policy review and evaluation）視域及方法而言，[2] 本人卻認為這份長達 174 頁（中文版）、經 35 位精英委員「辛勤工作」而成的《檢討報告》，存在明顯的不足與缺失。縱觀全份《檢討報告》，在沒有提供任何一般正規公共政策檢討和評鑑研究應有的政策效果實徵分析與論證的情況下，就「肯定系統評估的設立原意及價值」（教育局，2016a，頁 2），繼而就順理成章認定檢討工作只需要在實施細節（例如「試卷及題目設計」和「行政安排及報告」）上加以調整，就可完成整個檢討工作。從委員會成立之初就只分設「行政安排及報告檢討工作小組」和「試卷及題目設計檢討工作小組」，已可預見檢討工作只可能是在技術層面上的修補，而不會觸及系統評估的整體運作及方向的改革，更不會涉及系統評估的存廢問題。

從以上一種檢討的邏輯及實際工作的安排來看，檢討結果就必然是：前提決定了結論，即既然「肯定系統評估的設立原意及價值」，自然就只需稍事調整實施細節後，就可繼續「於 2017 年在全港實施」（教育局，2016a，頁 3）。然而，就公共政策研究領域中有關政策檢討和評鑑的視域及方法而言，本人必須提出以下連串的研究問題，向「工作組」各委員質詢：

1. 「系統評估的設立原意及價值」是甚麼？是否就如《檢討報告》所聲稱那般是「回饋學與教」（教育局，2016a，頁 2）呢？簡言之，檢討的第一個研究問題應是：清楚界定檢討政策的目標與意義。
2. 有關的政策意義與價值是否值得認同及肯定呢？簡言之，檢討的第二個研究問題應是：評定政策目標的可欲及可值性（desirability and worthiness）。
3. 有關的政策意義與價值是否已經得以實現？簡言之，實施了十年的系統評估是否已經實現了「回饋學與教」的政策目標？即檢討的第三個研究問題應是：政策措施的效能（effectiveness）問題。
4. 即使有關的政策目標確能通過相關的政策措施而得以實現，還必須追問整個政策實施過程付出了甚麼以至多少的代價及成本？即檢討的第四個研究問題是：政策措施的效率（efficiency）問題。
5. 政策實施過程中是否衍生一些未曾預計的後果？更甚者是：這些後果是否衍生出損害了政策原意的效果（self-defecting effect），以至對整體政策環境及生態（即學校教育生態）造成破壞性的後果？即檢討的最後一個研究問題就是：政策措施衍生的後果以至惡果。

對照以上政策檢討和評鑑的一系列基本研究問題，不難發現《檢討報告》只對首兩個研究問題在未經任何論證的情況下，就想當然地肯定了；至於其他三個更根本的研究問題，就完全隻字未提，更遑論進行研究及論證。更令人驚訝的就是，《檢討

報告》竟可以斷然作出結論：經微調後的系統評估即可繼續「於 2017 年在全港實施」（教育局，2016a，頁 3）。

對於一個實施了十年，每年影響小三、小六及中三近十萬名莘莘學子的學習以至福祉的教育政策，並受到數萬網民聲討及要求廢除的評估工具，教育當局卻可以接納這樣一個完全未有觸及政策檢討研究應處理的核心議題的《檢討報告》，並決定一年後就重新全面實施，本人實無法接受。據此，本人將在下文從三方面對系統評估這個影響深遠的教育政策重新檢討，它們分別是：（1）系統評估原意與價值的再檢討，（2）系統評估成效與後果的再檢討，（3）系統評估改革建議的再檢討。

系統評估原意與價值的再檢討：進展與總結性評估的辯證

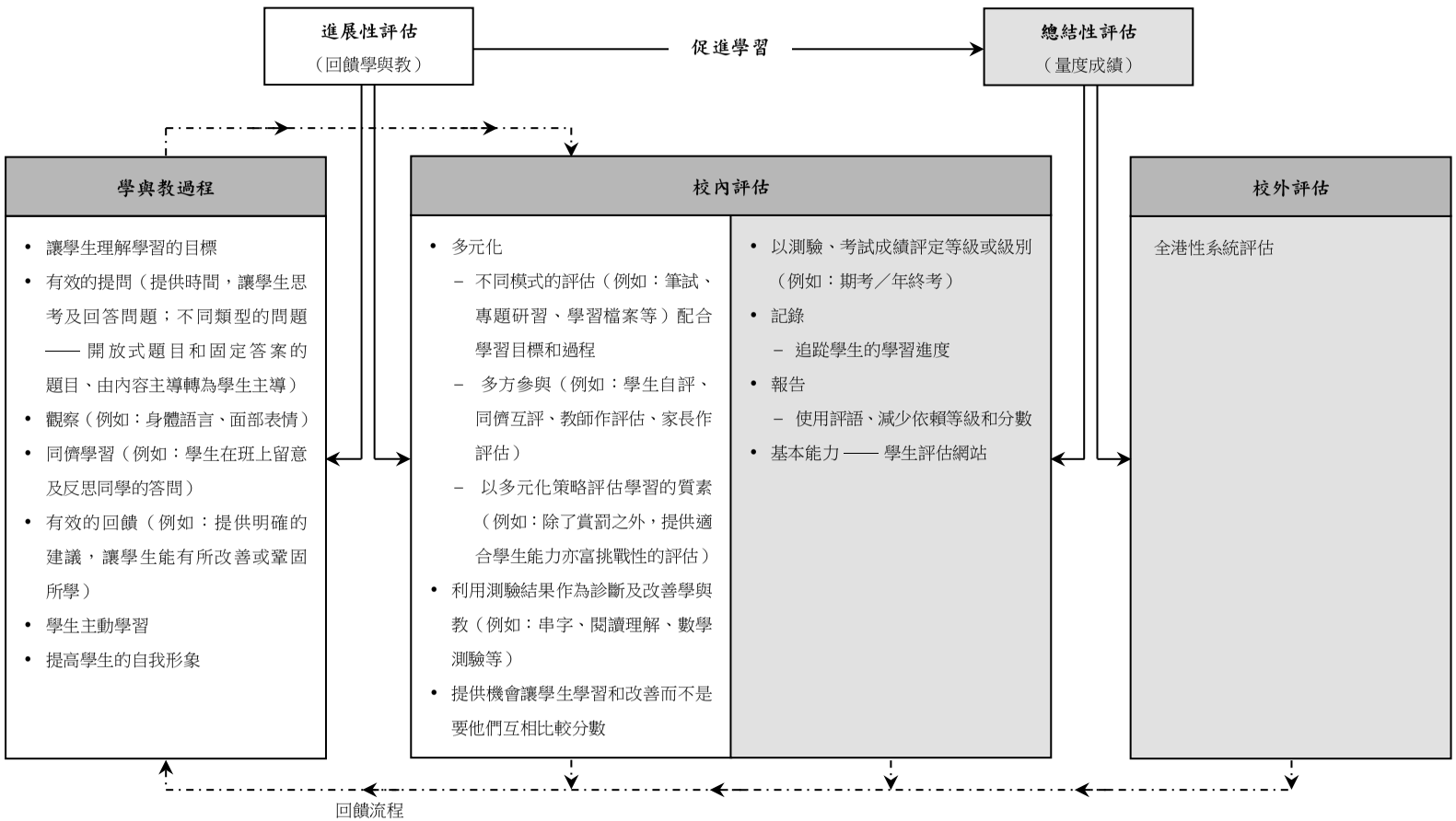
整份《檢討報告》的論證邏輯是建基在以下一個大前提：「肯定系統評估的設立原意及價值，認同系統評估提供的資料具有回饋學與教的功能」（教育局，2016a，頁 2）。據此，就建議系統評估只需要在技術細節層面（如「試卷及題目設計」和「行政安排及報告」）稍作調整，即可以重新全面「於 2017 年在全港實施」（頁 3）。然而，問題就在於：「系統評估的設立原意及價值」是否就真的為了「回饋學與教」呢？《檢討報告》對整份報告的前提與基礎完全未有加以論證，就想當然地作出了「肯定」與「認同」，以下就是本人對這個前提的真實性的質疑。

若要探討及闡釋一個已實施了十年的教育政策在制定及設計時的原來意義，在公共政策研究普遍運用的方法就是，回到當初的政策「文本」（text）及它在制定時所處的政策「議論」（discourse）的脈絡背景作分析。[3]

就政策文本而言，系統評估的設立源自本世紀初香港特別行政區（下稱特區）政府所公布關於教育改革及課程改革的兩份文件：《終身學習，全人發展：香港教育制度改革建議》（教育統籌委員會，2000）與《學會學習：課程發展路向》（課程發展議會，2001）；但兩份文件均只勾劃出評核改革應從傳統的成績量度的評估，轉向以促進學與教的進展評估方向進行改革，卻沒有概念化及操作化的具體說明。而有關的具體說明則可見於教育局網頁中一份名為《促進學習的評估》文件（教育局，2016b），文件開首的「學校實施評估的理念架構」（見圖一）就對系統評估作了清晰的定位。這個理念架構劃分評估為兩大範疇：

1. 「進展性評估」（formative assessment），並界定其功能是「回饋學與教」（informs learning and teaching）；
2. 「總結性評估」（summative assessment），並界定其功能是「量度成績」（measures attainment）。

圖一：學校實施評估的理念架構



資料來源：教育局（2016b）。

更重要的是，在架構內系統評估是放置在「總結性評估」的範圍下（見圖一最右方）；換言之，在評估改革的理念架構內，系統評估的首要功能就不是「進展性評估」的「回饋學與教」，而是「量度成績」。就此，本人就必須提出質疑：《檢討報告》堅稱系統評估的「設立原意及價值」是「回饋學與教」，卻隻字不提系統評估更主要的「量度成績」功能，是否有偷換概念以至指鹿為馬的嫌疑？

而且，在教育評估研究——特別是近三十年英、美有關教育改革、課程改革以至評估改革的文獻中，「進展性評估」與「總結性評估」之間的區分以至爭論至為明顯。例如，在英國的評估改革中，其中一位主導人物 Paul J. Black（倫敦 King's College 科學教育及教育評估教授、英國全國性評估改革工作組〔Task Group on Assessment and Testing, TGAT〕主席），就對「進展性評估」與「總結性評估」作出了清晰的區分；[4] 他更指出，TGAT 報告（1988）原先所設計及提議的一種「進展性評估」為本的評估改革，它結果之所以失敗告終，就是歸因於「總結性評估」的妨礙與壓制，[5] 即一種強調「量度成績」、製造排行榜、追求可量化的卓越指標的評估政策。這種強調追求簡單直接地量度成績的「總結性評估」，在上世紀末英國評估改革過程中，就侵佔和取代了另一種着重由教師主導、以課堂教學為本、長期及持續的「回饋學與教」的「進展性評估」取向。背後造成「總結性評估」對「進展性評估」殖民化的政策及政治動力就是，戴卓爾夫人所領導保守黨政府的新自由主義政治鼓吹的追求提升全球競爭力、增加公共部門（包括學校教育部門）的問責性及競爭性；其中對學校教育部門的改革，一方面就是力求加強對學校教育的監控（scrutiny）、審核（auditing）、問責（accountability）與懲治（sanction），另一方面則引入市場化機制，以增加家長的選擇。[6] 在這種政策大方向下，「總結性評估」所提供的量化且可比較的成績評量，就正好提供方便快捷的管治工具及市場信息。[7] 結果，英國上世紀末的評估改革就成為學校排行榜（school league table）、表現榜（performance table）及學生級別化的管治工具的犧牲品。據此，Paul J. Black 教授在回顧 TGAT 的改革經歷時，才會慨嘆是個「喪失了的機會」（opportunity lost）（Black, 1998）。

此外，美國在本世紀初由小布殊政府推行的教育改革，一般慣稱 *No Child Left Behind* 法例（原名是 Reauthorization of the Elementary and Secondary Education Act 2002），亦實施每年對全國三至八年級學生進行標準化評估，並據此向全國學校問責並加以懲治。對於這種高風險、以監控為主的「總結性評估」，不少學者（例如著名教育評估學者 W. James Popham）批評它與回饋及支援教學的「進展性評估」相違背，據此他再三呼籲，必須把美國的評估改革從注重監控與懲治的高風險「總結性評估」取向，轉移到支援教學的「進展性評估」取向上。[8]

若把以上英、美在評估改革中「進展性評估」與「總結性評估」之間的爭議與角力，用以審視本港系統評估這個評估工具的出現，我們就必須追問：系統評估制訂

與設立的原意是否亦像英、美教育改革那樣，是在新自由主義政治主宰下，為提升全球競爭力而對學校教育部門實施的一種監控、審核與問責的改革手段？若要回答這個問題，就有必要從當初系統評估制訂的政策議論脈絡入手，以理解系統評估的設立原意。

2004 年開始實施的系統評估，明顯與回歸後特區政府所鼓吹「共創香港新紀元」（1997 年特區政府首份施政報告的封面主題）的政策議論與願景有密切關係；這可證之於上述 2000 年及 2001 年出台的教育改革與課程改革文件所共用的頁面標題：「廿一世紀教育藍圖」；其次亦可證之於充斥於兩份文件的種種「政策巧語」（policy rhetoric），例如「知識型經濟」（knowledge economy）、「全球競爭力」（global competitiveness）、「提升整體質素」（improve the overall quality）等。若要實現這連串宏大願景，其中一個教育改革的原則（改革五大原則之一）就是：「講求質素」（教育統籌委員會，2000，頁 33）；據此，設立有效的「質素保證機制」自然成為不可或缺的改革工具，當中種種校外評估及監控機器更屬必不可少，這可證之於過去十年陸續強加於學校教學運作的一系列借鑑於商界核數機制的審核措施，如「質素保證視學」、「校外評核」、「學校自評」、「學校增值指標」及當然不可缺少的系統評估了。

至此，就更能理解為甚麼教育當局最初堅持全體學生均須參與系統評估，其目的是利用有關的成績量度，作每所學校「業績」審計的根據；系統評估這種評估措施實與英國的全國性評估和學校排行榜及表現榜的製作，或美國的全國標準測試及對「不及格」（failing）學校的懲治，本質上同出一轍。[9]

總結而言，本節運用政策文本及政策議論研究的方法，質疑《檢討報告》把「系統評估的設立原意及價值」認定為「回饋學與教的功能」及其背後的「進展性評估」並不真確。相反，本人的闡釋則是，系統評估的本質應是「總結性評估」及「量度成績」功能，而系統評估這種對全港中、小學童以至學校的「成績量度」更是整個學校質素保證及監控機制其中一個有力管治工具，這個機制是提高特區全球競爭力，以應付 21 世紀知識型經濟的教育改革議論的一個重要組成部分。

當然，以上對系統評估設立原意及價值的質疑與再界定，仍只停留在政策意義的闡釋。[10] 若要對《檢討報告》作更具體的批判，就有必要進一步從實徵層面入手，分析系統評估政策的成效與後果。

系統評估成效與後果的再檢討：促進學習與監控問責的辯證

過去一個世紀，公共政策研究領域中一個定論就是：一項公共政策的存廢，不是取決於它的設計原意及所欲價值，更不應以長官意志或當權者好惡為依歸，而應以

實徵研究驗證其成效為根據，然後把其成效對照於其成本甚至衍生的後果與損害，來決定政策的存廢。[11] 然而，《檢討報告》只根據「肯定系統評估的設立原意及價值」（教育局，2016a，頁 2），就決定重新全面「於 2017 年在全港實施」（頁 3）。縱觀 174 頁的報告，都沒有任何實徵研究數據，足以證立這個實施了十年，每年要求全體小三、小六及中三近十萬名學童應考的評估政策的成效；更具體而言，亦看不到任何因系統評估聲稱「具有回饋學與教的功能」而令本港學童「基本能力」有顯著提升的證據。

雖然《檢討報告》沒有提供任何實徵數據以證明系統評估政策的成效，但不代表有關數據不存在。事實上，從香港考試及評核局（下稱考評局）每年發布的《全港性系統評估學生基本能力報告》（下稱《系統評估報告》），就可以找到有關數據（見表一）。

表一：小三至小六學生系統評估成績的追蹤群組間分布，2004–2015（%）

	2004–2007	2005–2008	2007–2010	2008–2011	2010–2013	2012–2015
中國語文科						
小三及小六均達標	72.5	74.0	75.3	75.1	75.9	75.9
小三達標、小六不達標	11.9	11.9	11.4	12.0	11.4	11.1
小三不達標、小六達標	5.1	3.8	3.4	2.8	2.8	2.3
小三及小六均不達標	10.5	10.3	9.9	10.0	9.9	10.7
英國語文科						
小三及小六均達標	68.0	67.9	70.2	69.5	69.8	70.1
小三達標、小六不達標	9.6	11.3	10.8	11.0	10.2	10.0
小三不達標、小六達標	5.1	4.7	3.1	3.8	3.8	2.8
小三及小六均不達標	17.3	16.1	15.9	15.7	16.2	17.1
數學科						
小三及小六均達標	80.4	82.0	81.7	81.5	81.4	81.4
小三達標、小六不達標	6.3	6.8	6.7	7.0	6.9	6.7
小三不達標、小六達標	4.7	3.7	3.5	3.5	3.5	3.2
小三及小六均不達標	8.6	7.5	8.1	8.0	8.2	8.7

資料來源：考評局（2007，2008，2010，2011，2013，2015）。

在考評局每年發布的《系統評估報告》中，對參與系統評估的小學生作了以下一種追蹤的群組分類，它們分別是：[12]

1. 小三及小六成績均達標；
2. 小三達標、小六不達標；
3. 小三不達標、小六達標；
4. 由此亦可推論剩餘的一個群組，即小三及小六均不達標。

從表一所展示過去已有的六屆系統評估裏中、英、數達標率分布可以得知，系統評估一直未有發揮它聲稱具備的「回饋學與教的功能」和支援學習的「補底」功能。首先，這可證之於「小三及小六均不達標」學童群組的百分率，在過去六屆三個學科的有關百分率一直沒有改善。譬如以中國語文科為例，那些在小三通過系統評估已被發現未達標的約 10% 學童，若系統評估真正發揮其聲稱的「回饋學與教」功能，則部分學童應可在繼後小四、小五的學與教中，得到支援以至改進；但實際上在過去十年已有的六屆數據中，有關百分率一直沒有下降，始終穩定地維持在 10%（英國語文科的相關百分率是 16%，數學科是 8%）。據此可以推論，系統評估在小學階段沒有發揮它聲稱具備在學與教方面的「回饋」、「進展」以至支援及提升的功效。其次，考察表一中「小三不達標、小六達標」這學生群組，若系統評估真正具備「回饋學與教的進展性評估」的功效，預期這學生群組的百分率在過去十年會有所上升，但可惜有關百分率在過去六屆的三個學科均不斷下滑；以中國語文科為例，有關百分率就由 2004–2007 年的 5.1% 不斷下降至 2012–2015 年的 2.3%（英國語文科的相關數字是由 5.1% 下降至 2.8%，數學科是由 4.7% 下降至 3.2%）。換言之，若系統評估真的具備在學與教方面的「回饋」、「進展」與「補底」效能，這方面的效能十分微弱（三科均只對 5% 以下的學童產生效用），而且這方面的效能更是每況愈下。

事實上，若我們進一步審視考評局過去十年對系統評估成績的處理和發放方式及流程，就會發現整個流程以「總結性」（summative）的成績量度與發放為主，而且這些成績更只停留在全港系統及個別學校層面的量度，即包括各級與各科的全港系統性達標率、個別學校在各科的達標率、個別學校在各科試卷中個別考題的答對比率分布等。這樣總體性的成績量度，根本無法落實及下放到課堂層面的教學，以至個別學生的學習層面。具體而言，表一中顯示了那些在小三已被「診斷」出在特定學科未達標，甚至是學科內哪一能力範疇未達標的學童，由於發放到學校的數據停留在一種總體的形式（gross format），根本無法在課堂的學與教層面認定出這些學生來，他們亦無法得到適切的補救性教學支援（remedial instructional supports），更不能根本地力求使他們在學習活動與模式中產生「進展性的改變」（formative changes in students' learning activities and patterns）。[13]

考評局對系統評估成績處理工作上的另一個缺失，是成績發放方式只停留在「回顧」（retrospective）的「試後檢討」層面。據了解，在為教師提供的相關研討會及工作坊，討論內容主要集中在當年的系統評估成績表現，以及各級各學科以至各考題的表現，甚少作「進展」（formative）、「前瞻」（perspective）、「發展」（progressive）的教學支援的討論。在這樣一種「賽後檢討」的發放方式以至思維模式下，前線教師自然容易被引領進入一種「應試教學」的工作模式內，即努力為下一屆系統評估的

應考學生作準備，使他們不致在相類近的考題上犯上相同錯誤；而並非為已應考但不達標的學生進行「補救性」及「進展性」的教學支援。[14]

根據以上對六屆小學生在三個學科的系統評估成績表現的追蹤分析，加上對考評局處理與發放系統評估成績方式的考察，本人有理由相信過去十年系統評估的實施，根本沒有實現「回饋學與教的進展性評估」的功效。從考評局發放的系統評估成績的形式來看，系統評估只停留在「量度成績的總結性評估」的功能而已。

若我們把系統評估這種「量度成績的總結性評估」功能，放置回到過去十年特區政府教育當局努力建構的整個「教育質素保證」的政策議論脈絡內，特別是連結到其他的教育質素監控、審核及問責的政策工具，例如學校表現指標、質素保證視學、學校外評及自評機制、學校增值指標、學校表現評量 23 條等，就自然明白系統評估過去十年真正的政策效能，就是提供一套有力的「量度成績的總結性評估」工具，並把全港中、小學網羅在它的監控視野之內，以進行年度「業績」的審核以至問責。

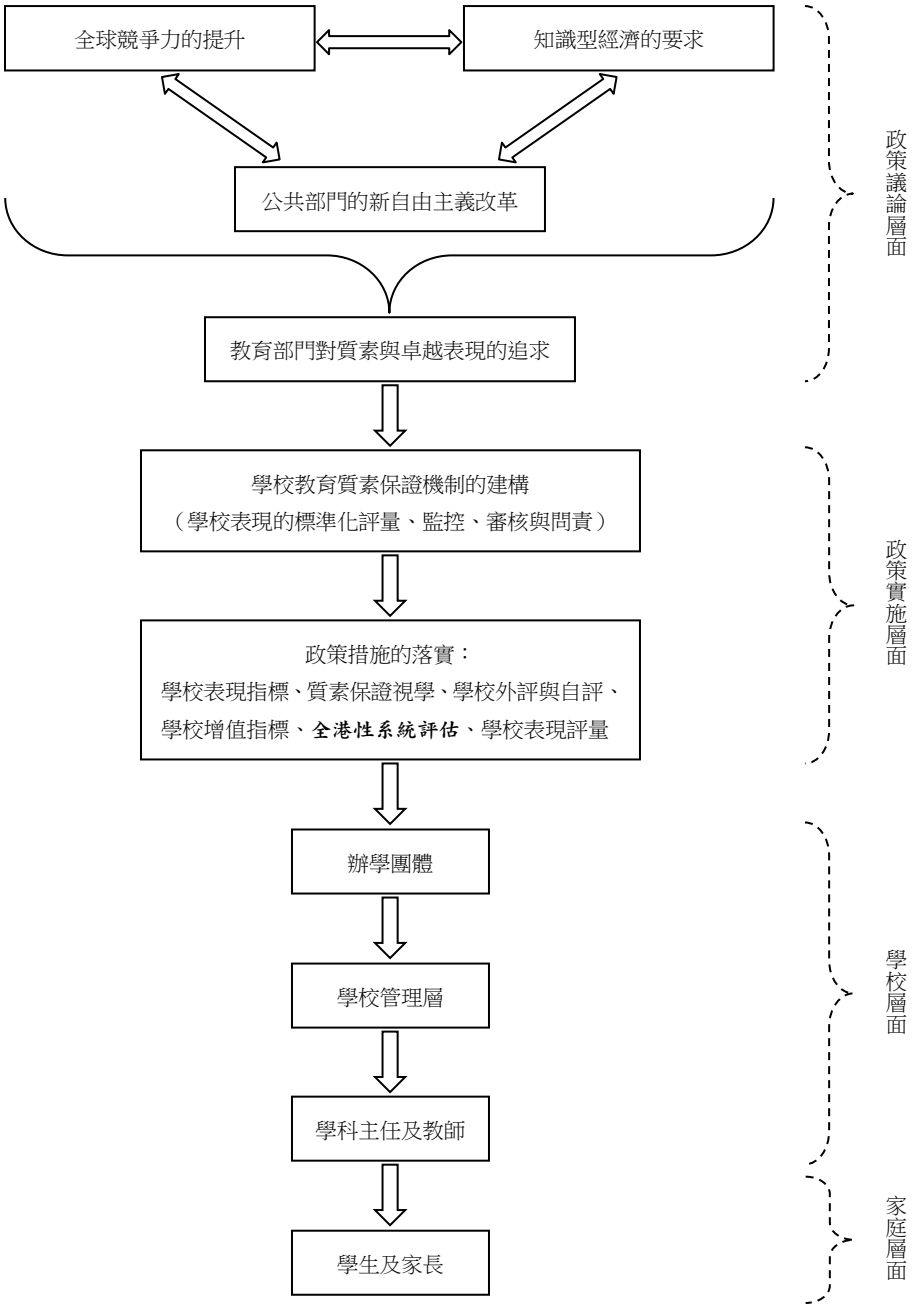
亦是在這樣一種監控、審核及問責的政策效能下，系統評估就為全港中、小學帶來年復一年的沉重壓力。在是次系統評估爭議中，一位資深小學校長的下述一段公開言論，就最能說明系統評估在這方面的政策後果：

> TSA〔全港性系統評估〕的「達標率」這個學校表現指標，原本是教育局外評隊手上的一把尺，但在學校看來就等於一把架在頸上的利刃……令學校內部承受不少的壓力。（張勇邦，2015）

除了校長和教師以外，在整個學校質素監控、審核及問責的政策「傳遞鏈」（delivery chain）[15] 中，生活在最底層的學生及他們的家長，所承受的壓力就更明顯。這可證之於 2015 年底在短短兩個月內，在網上即可發動五萬多人支持「取消小三系統評估」的群眾運動。[16] 事實上，最能說明系統評估在質素監控、審核與問責的政策議論下，所引發及造成對整個學校教育體系的一種操練壓力的理念架構，就是 Stephen J. Ball 及其同事對英國「全國性評估」的政策「傳遞鏈」的批判。圖二引用 Ball 等人的理念架構，來說明系統評估對香港學校教育體系所造成的操練壓力的因果關係「傳遞鏈」（Ball, Maguire, Braun, Perryman, & Hoskins, 2012, p. 515, Fig. 1）。

在上述一種系統評估所造成的監控、審核及問責的政策效能，以及導致沉重壓力的政策後果下，教育局高官及他們的學界辯護士卻仍辯解系統評估是低風險及與操練無關的政策時，才會觸發更巨大的輿論壓力，致使教育局要急急成立「工作組」以回應社會訴求。據此，必須進一步審察《檢討報告》中的改革建議是否能對症下藥，以消解系統評估在過去十年所造成的政策效能及後果。

圖二：質素保證政策的傳遞鏈 —— 追求表現壓力的層壓式傳遞



系統評估改革建議的再檢討：促進學習評估的回歸

根據上文第一節政策文本與政策議論的分析，本人質疑《檢討報告》把系統評估的「設立原意及價值」認定為「回饋學與教」的「進展性評估」是否真確，並提出

系統評估「設立原意」應屬「量度成績的總結性評估」，繼而更論證系統評估實屬特區政府過去十年在努力建構的一套教育質素監控及問責機制重要組件之一。又根據上文第二節政策效能與後果的分析，本人提出過去十年系統評估的實際效果，並未見實現了「回饋學與教的進展性評估」的效能；相反，論證了過去十年系統評估卻發揮着一種監控及問責的高風險「量度成績的總結性評估」的政策效能，並對學校校長、教師、家長及學生帶來沉重壓力。就以上系統評估的政策現況分析，本人有理由質疑《檢討報告》中提出的政策改革建議，是否能對症下藥？以下是本人的論證。

首先，本人認為系統評估爭議的解決方案必須回到本世紀初教育評估改革中兩大範疇之間的區別，即「進展性評估」與「總結性評估」。雖然概念上「進展性評估」與「總結性評估」未必互相排斥，但當把它們置於上世紀末的教育改革潮流，特別是英、美那種因全球化競爭壓力所湧現的新自由主義政治思維下，評估改革就從原先的「進展性評估」取向逐步為「總結性評估」所騎劫以至取代；亦即一種強調學生學習為本、教師參與、學與教互動的評估取向，就在強調學校教育質素監控與問責的政策議論壓力下，逐步變質為一種強調外控、標準化量度成績、附帶問責後果的高風險「總結性評估」。其中本文第一節提及，英國的評估改革被「學校排行榜」與「表現榜」所支配，就最能說明英國評估改革中「進展性評估」與「總結性評估」之間的此消彼長，只能變為「喪失了的機會」（Black, 1998），這彰顯了評估改革歷程中的核心問題所在。

在香港的評估改革過程與爭論中，雖然在原初改革文件中（見圖一）仍看到兩大範疇的區分與識別，但繼後的發展與爭論，在官方的「政策詞彙」（policy terminology）中，系統評估的「總結性評估」功能及其衍生的政策效能與後果卻完全「被失蹤」；甚至當家長及教育界在網上發出對系統評估的強烈聲討時，教育局的高官及其學界辯護士竟可公開聲稱系統評估屬低風險、且與應試操練壓力無關。[17] 這種態度若非顯示他們對系統評估本質的無知，就是他們根本在掩耳盜鈴。

更甚者，在《檢討報告》中，我們仍找不到對系統評估的「總結性評估」本質、其監控及問責的政策效能，及其衍生的高風險應試操練壓力的政策後果有所正視和處理的建議。報告更把高風險應試操練壓力推諉成為只是「不同持份者對評估風險的理解」上的差異（教育局，2016a，頁 6），因此只把問題界定為「溝通」與認知的問題，解決方法就只是「加強公眾教育，提升各界的評估素養」（頁 7）。

然而，必須強調，系統評估的「總結性評估」本質所衍生的高風險應試操練壓力是真實存在的現象，而非只屬主觀的「理解」與「素養」問題。因此解決方案必須針對造成高風險應試操練壓力的根源，即系統評估對學校教育構成的那些質素監控、審核與問責的政策後果。出奇的是，雖然《檢討報告》仍繼續不承認系統評估所造成的高風險應試操練壓力與後果，例如報告繼續聲稱「教育局向學界重申系統評估屬

低風險的評估」（教育局，2016a，頁 6），但卻「從善如流」地提出：「從 2016/17 學年起，把系統評估從『學校表現指標』中『8.1 學業表現』的要點問題中移除」（頁 6）。據此，本人希望教育當局能真誠確實地把系統評估從整個教育質素保證機制中割離出來，而機制中其他政策工具（例如「學校表現評量」23 條）亦一併從系統評估中移除，並敦促前線外評視學人員切實執行上述建議。簡言之，請把「架在學校頸上的利刃」拿走。

若系統評估的「總結性評估」功能不再用於監控、審核與問責個別學校的表現，則系統評估的實施方式便不必覆蓋全港學校的所有學生，因為當系統評估的「總結性評估」功能回到其真正的設立原意，在「全港性系統」層面量度學校教育制度的表現，則藉科學化的隨機抽樣而選取具代表性的學校樣本，已可提供有效且可靠的「全港性系統評估」了。換言之，當系統評估的「總結性評估」只是提供教育界一種全港參考指標，而不再是強加於每所學校的監控、審核與問責的管治工具，則更無法理解為何《檢討報告》還要堅持在一年後就重新強迫全港所有學校的小三、小六及中三近十萬名學童應考系統評估。

若系統評估真正能與教育質素保證機制完全脫勾，並只以隨機抽樣方式進行，本人相信系統評估有可能擺脫高風險應試操練的詛咒，而回歸到「基本能力評估」（Basic Competence Assessment，下稱 BCA）這個更大的評估改革框架內，並切實擔當起其「總結性評估」為全港提供系統性表現的參考指標。據此，個別中、小學在系統評估中的參與就不再是過往十年那種外控、強制的方式，而可以是一種自發（self-initiated）、校本（school-based）、分散（decentralized）的參與形式。具體而言，對每年不在抽樣名單內的大部分學校，它們仍然可在每年系統評估進行抽樣評估後，自行採用當年已公開的系統評估試卷，對校內學生進行評估，並把本校成績對照於公布的系統評估成績，便可自行量度校內學生的基本能力達標率，甚至可以分析校內學生在各特定能力範疇的強弱水準。換言之，各學校均可自行製作類似現行由考評局發放的系統評估學校報告。更重要的是，每所學校更可以自行把校內成績仔細分析至特定班別以至個別學生層面，因而能真正做到把系統評估的總結性成績量度「回饋」到課堂教學以至學生學習，即能給予學生適時的基本能力診斷及補救性的教學支援。而且，這種校本、自發及分散的系統評估，更可避免校方、學童以至家長對外來公開考核產生恐懼與壓力，甚至更能真正實現一種促進學習而又人本的評估改革。

事實上，過去十年，整個 BCA 的改革中，系統評估其實只是其中一個輔助性的「總結性評估」工具，但由於學校質素監控、審核與問責的政策議論的支配，致令系統評估在 BCA 中提供系統性參考指標的功能未能彰顯，反而衍生種種應試操練的壓力。事實上，《檢討報告》不斷強調的「回饋學與教的功能」其實是屬於 BCA 框架中「進展性評估」內的「學生評估」（student assessment）部分，即一個網上評估資源

庫及作業系統，其中個別學校及教師是可以自發而主動參與及運用該系統。事實上，這部分的 BCA 才是整個評估改革的焦點與重心。

「進展性評估」為本的評估改革，重點應是發揮課堂教師教學及學生學習的積極性，使他們能有效運用 BCA 作回饋、協助及促進學生學習的工具，而不應該是為政府內的技術官僚提供量度前線業績有力的問責與管治工具。寄望是次系統評估的爭議與檢討，能把評估改革回歸到促進莘莘學子學習與福祉的基礎上，而不再是技術官僚追求管治效率的手段。

註 釋

1. 見《文匯報》2016 年 2 月 5 日 A4 版報導。
2. 有關討論可參考 Fischer（1995）、Hudson & Lowe（2004）、Nagel（2002）、Pawson（2013）、Rist（1995）。
3. 有關在公共政策研究上的「文本－議論」取向，可參考：Ball（1994）、Fischer（2003）；亦見曾榮光（2007）。
4. 可參考 Black（1992, 1993, 2000）和 Black, Harrison, Lee, Marshall, & Wiliam（2002）。當中「進展性評估」被界定為「所有教師及／或學生從事的活動，其目的是提供信息，得以回饋用以改善這些師生正在從事的教與學活動」（Black & Wiliam, 1998, pp. 7–8）。據此，Black et al.（2002）更列出教育評估的三項主要功能：（i）改善教學（improving instruction）與促進學習（promoting learning），（2）問責（holding accountable）於教學單位（包括學校及教師），（3）選拔與驗證（selecting and certifying）學生；並強調「進展性評估」的首要優先序是促進學生學習，而非服務於問責（accountability）或對學生能力的排等或驗證。至於「總結性評估」則主要是在學習完結階段，用作問責及／或驗證的評估；他更強調「總結性評估」「雖然重要，但亦可以對教與學造成重大的破壞」（do great damage to the business of teaching and learning）（Black, 1992, p. 3）。此外，美國評估學者 James Popham 對「進展性評估」與「總結性評估」亦作出了類似區分：前者是提供「由評估而取得的證據」（assessment-elicited evidence），用作「改善失敗但仍可改進的教學」（to improve unsuccessful yet still modifiable instruction）；相反，後者的目的是作出「最終成／敗的決定」（final successful/failure decision），並加之於一些「相對地已無從改進的教學活動」（relatively unmodifiable instructional activities）（Popham, 2011, p. 10）。
5. 可參考 Black（1997, 1998）；亦可參考 Lawton（1992）和 Murphy（1990）。英國的評估改革原意是把着重「總結性評估」轉變為以「進展性評估」為本的政策改革（Black, 1998; Lawton, 1992），但由於戴卓爾政府偏向於新自由主義的市場化改革，並為求學校教育市場化及家長擇校得以可能而追求一種簡單、快捷且可靠的「市場信息」，致使「進展性評估」只能淪為「總結性評估」的附庸，甚至徹底被異化（Black, 1998; Murphy, 1990）。

6. 有關把學校教育視作公共部門的問責性改革（accountability reform），可參考：Ball（2008）、Chitty（2014）、Clarke, Gewirtz, & McLaughlin（2000）、Gleeson & Husbands（2001）。
7. 有關「全國性評估」（National Assessment）與市場信息的關係分析，見 Murphy（1990）；亦可參考 Ball（2008）。
8. 可參考 Popham（2001, 2003, 2004）；亦可參考 Carnoy, Elmore, & Siskin（2003）和 Sleeter（2007）。
9. 有關香港特區質素保證機制的政策發展，以及與英、美相關政策的比較討論，可參考曾榮光（2011a）。相較於英、美質素保證機制（包括質素的標準化評量、監控、審核、問責與懲治，香港的有關政策措施在程度上仍未推展到懲治（例如收回辦學權、關閉學校等）的階段，可說只推展到問責階段為止。
10. 有關公共政策的意義闡釋研究，可參考曾榮光（2011b）、Majone（1989）、Wagenaar（2011）。
11. 有關政策分析以至政策科學的發展，討論眾多，例如 deLeon（1997, 2006）、deLeon & Martell（2006）、Fischer, Miller, & Sidney（2007）、Parsons（1995）。
12. 有關數據的報導，最初見「香港 01」記者何敬淘（2016）的網上報導。
13. 根據 Black & Wiliam（2009），評估所提供的回饋信息可稱得上是「進展性主要視乎有關學生成績的證據是否得教師、學習者及其同儕所採納、闡釋及運用，以作決定下一階段改善教學的依據」（p. 9）。據此可以論定，過去十年考評局發放到學校的所謂回饋信息就稱不上是「進展性」，因為有關學生成績的證據只停留在系統與學校層面，從來無法下放到 Black & Wiliam（2009）所界定「進展性評估」的三個層次：教師的課堂教學、學生同儕之間的互動、個別學生的學習改善（pp. 7–8）。此外，若回歸到本文第一節提及教育局採用的「學校實施評估的理念架構」，亦不難從架構中「進展性評估」範疇下（見圖一左方）找到上述提及的「課堂教學」、「同儕學習」與「學生學習」的三個層次，同時如「診斷」、「改善」等「進展性」的步驟。這在在說明過去考評局向學校發放的所謂「回饋學與教」的信息，根本並非「進展性評估」概念所界定的「回饋」。
14. 有關「進展性評估」的學與教實踐，可參考 Black et al.（2003）、Heritage（2013）、Popham（2003）。
15 「傳遞鏈」是英國教育家及國際企業顧問 Michael Barber 的概念（見 Barber, 2007）。他曾任英國教育部有關學校標準的首席顧問，並曾任香港特區教育局顧問。有關英國全國性課程與評估的政策傳遞鏈對教師與學生及家長所造成的壓力，見 Ball, Maguire, Braun, Perryman, & Hoskins, 2012。
16. 可參考 https://www.facebook.com/events/525241554300641/
17. 有關討論及批評，可參考曾榮光（2015a，2015b，2016）。

參考文獻

何敬淘（2016）。〈吳克儉：TSA 改善教學　數據：應試小學生成績無改善〉。擷取自 http://www.hk01.com/港聞/3756/吳克儉-TSA 改善教學-數據-應試小學生成績無改善

香港考試及評核局（2007）。〈2007 年全港性系統評估學生基本能力報告〉。擷取自 http://www.bca.hkeaa.edu.hk/web/TSA/zh/2007tsaReport/priSubject_report_chi.html

香港考試及評核局（2008）。〈2008 年全港性系統評估學生基本能力報告〉。擷取自 http://www.bca.hkeaa.edu.hk/web/TSA/zh/2008tsaReport/priSubject_report_chi.html

香港考試及評核局（2010）。〈2010 年全港性系統評估學生基本能力報告〉。擷取自 http://www.bca.hkeaa.edu.hk/web/TSA/zh/2010tsaReport/priSubject_report_chi.html

香港考試及評核局（2011）。〈2011 年全港性系統評估學生基本能力報告〉。擷取自 http://www.bca.hkeaa.edu.hk/web/TSA/zh/2011tsaReport/priSubject_report_chi.html

香港考試及評核局（2013）。〈2013 年全港性系統評估學生基本能力報告〉。擷取自 http://www.bca.hkeaa.edu.hk/web/TSA/zh/2013tsaReport/priSubject_report_chi.html

香港考試及評核局（2015）。〈2015 年全港性系統評估學生基本能力報告〉。擷取自 http://www.bca.hkeaa.edu.hk/web/TSA/zh/2015tsaReport/priSubject_report_chi.html

張勇邦（2015）。〈津貼小學議會主席張勇邦 —— 如何改革 TSA？〉。擷取自 http://programme.rthk.hk/channel/radio/programme.php?name=radio1/hkletter&d=2015-10-24&p=1085&e=329850&m=episode

教育局（2016a）。《全港性系統評估檢討報告》。擷取自 http://www.edb.gov.hk/attachment/tc/curriculum-development/tsa/fullreport.pdf

教育局（2016b）。〈促進學習的評估〉。擷取自 http://www.edb.gov.hk/tc/curriculum-development/assessment/about-assessment/assesment-for-learning.html

教育統籌委員會（2000）。《終身學習，全人發展：香港教育制度改革建議》。香港，中國：教育統籌委員會。

曾榮光（2007）。〈教育政策研究：議論批判的視域〉。《北京大學教育評論》，第 5 卷第 4 期，頁 2–30。

曾榮光（2011a）。〈從教育質素到優質教育的議論〉。載曾榮光，《香港特區教育政策分析》（頁 195–219）。香港，中國：三聯書店；香港浸會大學當代中國研究所。

曾榮光（2011b）。〈理解教育政策的意義 —— 質性取向在政策研究中的定位〉。《北京大學教育評論》，第 9 卷第 1 期，頁 152–180、192。

曾榮光（2015a，12 月 23 日）。〈TSA 低風險？〉。《明報》，觀點版。擷取自 http://news.mingpao.com/pns/dailynews/web_tc/article/20151223/s00012/1450806930735

曾榮光（2015b，12 月 30 日）。〈TSA 有效？與操練無關？〉。《明報》，觀點版。擷取自 http://news.mingpao.com/pns/dailynews/web_tc/article/20151230/s00012/1451411975670

曾榮光（2016，1 月 15 日）。〈辨別 TSA 的真實意義：比較視域的分析〉。《明報》，觀點版。擷取自 http://news.mingpao.com/pns/dailynews/web_tc/article/20160115/s00012/1452793365024

課程發展議會（2001）。《學會學習：課程發展路向》。擷取自 http://www.edb.gov.hk/tc/curriculum-development/cs-curriculum-doc-report/wf-in-cur/

Ball, S. J. (1994). *Education reform: A critical and post-structural approach*. Buckingham, England: Open University Press.

Ball, S. J. (2008). *The education debate*. Bristol, England: The Policy Press.

Ball, S. J., Maguire, M., Braun, A., Perryman, J., & Hoskins, K. (2012). Assessment technologies in schools: "Deliverology" and the "play of dominations." *Research Papers in Education*, *27*(5), 513–533. doi: 10.1080/02671522.2010.550012

Barber, M. (2007). *An instruction to deliver: Tony Blair, the public services and the challenge of delivery*. London, England: Politico's.

Black, P. (1992). *Assessment of student learning: Relating UK experience to the Australian context*. Victoria, Australia: The Incorporated Association of Registered Teachers of Victoria.

Black, P. (1993). Formative and summative assessment by teachers. *Studies in Science Education*, *21*(1), 49–97. doi: 10.1080/03057269308560014

Black, P. (1997). Whatever happened to TGAT? In C. Cullingford (Ed.), *Assessment versus evaluation* (pp. 24–50). London, England: Cassell.

Black, P. (1998). Learning, league tables and national assessment: Opportunity lost or hope deferred? *Oxford Review of Education*, *24*(1), 57–68. doi: 10.1080/0305498980240105

Black, P. (2000). Research and the development of educational assessment. *Oxford Review of Education*, *26*(3&4), 407–419. doi: 10.1080/713688540

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom*. London, England: nferNelson.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, England: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 7–74. doi: 10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. doi: 10.1007/s11092-008-9068-5

Carnoy, M., Elmore, R., & Siskin, L. S. (Eds.). (2003). *The new accountability: High schools and high-stakes testing*. New York, NY: RoutledgeFalmer.

Chitty, C. (2014). *Education policy in Britain* (3rd ed.). New York, NY: Palgrave Macmillan.

Clarke, J., Gewirtz, S., & McLaughlin, E. (Eds.). (2000). *New managerialism, new welfare?* London, England: Sage.

DeLeon, P. (1997). *Democracy and the policy sciences*. New York, NY: State University of New York Press.

DeLeon, P. (2006). The historical roots of the field. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 39–57). Oxford, England: Oxford University Press.

DeLeon, P., & Martell, C. R. (2006). The policy sciences: Past, present, and future. In B. G. Peter & J. Pierre (Eds.), *Handbook of public policy* (pp. 31–47). London, England: Sage.

Fischer, F. (1995). *Evaluating public policy*. Belmont, CA: Wadsworth/Thomson Learning.

Fischer, F. (2003). *Reframing public policy: Discursive politics and deliberative practices*. Oxford, England: Oxford University Press.

Fischer, F., Miller, G. J., & Sidney, M. S. (Eds.). (2007). *Handbook of public policy analysis: Theory, politics, and methods*. Boca Raton, FL: CRC Press.

Gleeson, D., & Husbands, C. (Eds.). (2001). *The performing school: Managing, teaching and learning in a performance culture*. London, England: RoutledgeFalmer.

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action*. Cambridge, MA: Harvard Education Press.

Hudson, J., & Lowe, S. (2004). *Understanding the policy process: Analysing welfare policy and practice*. Bristol, England: The Policy Press.

Lawton, D. (1992). Whatever happened to the TGAT report? In C. Gipps (Ed.), *Developing assessment for the National Curriculum* (pp. 95–103). London, England: Kogan Page.

Majone, G. (1989). *Evidence, argument, and persuasion in the policy process*. New Haven, CT: Yale University Press.

Murphy, R. (1990). National assessment proposals: Analysing the debate. In M. Flude & M. Hammer (Eds.), *The Education Reform Act, 1988: Its origins and implications* (pp. 37–49). London, England: Falmer Press.

Nagel, S. S. (Ed.). (2002). *Handbook of public policy evaluation*. Thousand Oaks, CA: Sage.

Parsons, W. (1995). *Public policy: An introduction to the theory and practice of policy analysis*. Cheltenham, England: Edward Elgar.

Pawson, R. (2013). *The science of evaluation: A realist manifesto*. London, England: Sage.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2004). *America's "failing" schools: How parents and teachers can cope with* No Child Left Behind. New York, NY: RoutledgeFalmer.

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th ed.). Boston, MA: Pearson.

Rist, R. C. (Ed.). (1995). *Policy evaluation: Linking theory to practice*. Aldershot, England: Elgar.

Sleeter, C. E. (Ed.). (2007). *Facing accountability in education: Democracy and equity at risk*. New York, NY: Teachers College Press.

Wagenaar, H. (2011). *Meaning in action: Interpretation and dialogue in policy analysis*. Armonk, NY: M. E. Sharpe.

A Critique on the *Report on Review of the Territory-wide System Assessment*: In the Perspective of Policy-evaluation Research

Wing-Kwong TSANG

Abstract

On 29 February 2016, the Report on Review of the Territory-wide System Assessment *(TSA) was released on the official website of the Education Bureau of the HKSAR government. In this report, we cannot find any empirical analysis or justification on the effects and consequences of the policy in point, as we would expect of any formal policy-review study; yet we are told to accept its conclusion that the intent and value of the TSA is worth to be "reaffirmed" and what is needed is some tinkering on the implementation procedures. As a result, we are also told to accept the fact that the "improved" TSA will be launched in 2017 for "territory-wide implementation." In this article, the Report will be examined and critiqued from the perspective of policy-evaluation research. First, the article will query and critique the basic premise assumed by the Report, that is, "the intent and value of the establishment of TSA" is "to provide feedback to learning and teaching." Second, it will argue empirically that for the last ten years, the TSA policy has failed to serve as an effective formative-assessment instrument to provide relevant feedbacks and supports to learning and teaching of students, especially those who are lagging behind in the Basic Competence Assessment. On the contrary, the article will argue that the TSA policy has espoused a policy consequence of unleashing a wave of pressure of teaching for testing, which has been permeating widely among schools, especially primary schools in recent years. Finally, the article will take issue with the primary policy recommendation of the Report. It will argue that it is totally unnecessary to continue a "territory-wide" TSA to mandate all students of Primary 3, Primary 6 and Secondary 3 to sit for the assessment annually. The article will suggest that the coverage of the TSA can be scaled down to a randomly selected sample of schools, and argue that such a scaled-down TSA could still provide reliable and valid evidence for the school performance at system level.*

Keywords: Territory-wide System Assessment; formative assessment; summative assessment; quality-assurance mechanism; policy-evaluation research