目 錄

避免由火坑掉進深淵：
對升中能力測驗的一些建議

摘 要

教育委員會屬下之學校教育檢討小組，最近發表了九年強迫教育檢討報告，全面回顧中學學位分配的辦法，並建議以「語文及數學能力評估」（學科試）替代現行的「學能測驗」（學能試）。對小學以低劣練習過分操練學能試的情況，我們深感問題嚴重，但我們不同意用學科試導引小學課程，便可解決困難。我們認為並無理想試題類型，學科試只會迫使學校依一狹窄課程瘋狂地操練學生，無視其他更廣泛更重要的教學目標。

本文依據報告書的大前提，以更接近校內課程的考試取代學能試的原則下，回顧香港及其他國家的教育改革，提出一套方案，這包括矩陣取樣，每一學校的不同學生，可能接受多套內容不同的試卷測試，使考試內容及形式（包括口試），更為多樣化。此外，我們認為用於調整各校成績的考試（學科或學能試），應只擔當調整角色，並減低其對學校正常教學的干擾；對大部分學校，其產出水平（調整試成績）穩定，我們只需多年才評審一次。這些方法是統計上有效，且更為易於銜接日後學校本位考試、表現性評估、目標為本課程及其他教育改革。

導 言

教育委員會屬下之學校教育檢討小組，最近發表了九年強迫教育檢討報告（Board of Education, 1997a, 1997b；簡稱《報告書》），全面回顧中學學位分配的辦法，並建議以「語文及數學能力評估」（簡稱「學科試」）替代現行的「學能測驗」（簡稱「學能試」）。考試方法及內容直接影響課堂教學，生死攸關的考試影響最大；各國經驗如此，香港也無例外，也許更甚。

有些人指出今次檢討升中派位機制是由前線工作者（校長教師等）主動提出（Joint Committee on Review, 1996），由下而上的教育改革訴求，與其他由教育政策者主導的改革並不相同，這也足以說明「學能試」的負面影響在中小學前線工作者中已達不能再忍耐的沸點，如今很多建議，雖然深明可能導致其他問題，但在迫切追求改革，希望盡早取消學能試的形勢下，一些不太成熟且可能導致教育倒退的建議也推出市場，這雖然顯示學能試操練問題的嚴重性，但各類建議也引來激烈爭論。

本文主要是總結一些近期爭論要點，指出報告書內部分建議不足之處，在依據報告書的大前提，以更接近校內所教課程的考試取代學能試，提出一套方案，務使這只用於調整各校成績的考試，減低對學校正常教學的干擾，令一些致力優質教育的學校，有更大自主空間發展其理想教學，無需被迫依一狹窄課程操練一些無助真正學習的低劣練習；這調整學校間成績的機制也能更易銜接日後課業

式、表現性（performance test）、學校本位（school based）甚至 TOC 等所倡議的學業評估方法。

一、現行制度

1978 年以前，小六學生須參加一中、英、數的升中試，個人考試成績直接決定學生派位，頗多學校及家長，只注重這三主科，而忽略其他科目及活動，令小學課程未能均衡發展。1978 年後，學生是否獲派其所選學校，主要取決於其校內絕大部分科目的總成績，爲了調整各校的評分標準，所有學生需參加一公開學能測驗（簡稱「調整試」），內容爲文字及數字推理。不過學生的學能測驗得分，只用於學校間調整，而不計算在其個人派位名次內。爲減低考試之競爭性，及將不同能力的學生作某種程度之混合，各學生將獲電腦分配一隨機號碼，故學生派位次序是以其調整後的校內成績，加上運氣的因素而決定。

二、解決學能試操練的兩種思路

在二十年前所開始採用的升中派位機制中，我們用種種方法（如：採用學能試而非學科試，不公布試題）去減低「調整試」對正常課堂教學的影響，當初這制度可能運作甚好，現在有頗多學校正常教學受操練學能試所影響，要解決這困難有兩種截然不同的策略。

第一類（簡稱「調整試導引改革」）方法是若果無法消除操練，則改爲操練有用的內容，這是現在《報告書》所建議的策略，這包括改考一些與課

程有關的內容，公布試題以「指導」操練，這是以考試引導課程教學改革的思路。第二類（簡稱「調整試只作調整」）方法是採用種種方法，去減低操練的報酬，其中可包括「就近入學」及「完全隨機派位」（即將五等級改爲一等級，全由運氣決定）。

「調整試導引改革」的想法是改變調整試的內容，依現在的建議必導致提高其對教學的影響，「調整試只作調整」的想法是以種種措施，以達至更減低其對教學之影響；這是兩種截然對立的思路。我們將討論前者可能帶來的問題，並提出在改變調整試（改用較多學校課程有關之內容）的前提下，列舉一些可行措施及路向。

《報告書》及其引發之問題

一、對學能試的不公允批評

《報告書》臚列各項對學能試的批評，多不公允，未能道出這測驗的原來目的，且將人爲錯誤（如：過分錯誤操練），委於測驗欠佳。

學能試的唯一作用是調整各校成績，爲了令各校依其教學理想及學生程度，在教學內容及方法能有更大的自主權，故考試盡量不劃定範圍，亦不公布試題，免得正常教學受這測驗所左右。在有限的資源下，考試時間頗短，每科 45 分鐘，一個下午考畢。因該試不直接計入個別學生成績內，所以對學生壓力也甚低。

學能測驗與學科（校內）成績有極高的相關，也說明以此作各校調整指標甚爲適宜。這方法可讓各校有足夠專業自主教學，對學生壓力不大，且合理地調整各校成績，這在十多年前開始運作的設計實在充滿智慧。

《報告書》建議公布測試範圍，定期公開評估樣本試題，長遠而言，並應公開所有測試題目。這些措施並不恰當。首先，好的試題及類型甚難撰寫設計，故不宜公開。最重要的是一旦公開樣本，不難預見極多老師以這些樣本（或其他類似坊間作業），作大量不依教學課程的操練。因考試時間甚短（現在只得 45 分鐘），試題種類有限，我們極不願見到老師，以這三幾類試題作他們教學的指揮棒。

對於一些相對較難受操練影響的學能／性向測驗，已發覺頗多教師作過度操練，不難預見，若改爲更可以操練的語文／數學能力評估，教師必然針對數類樣本試題作更密集的操練，影響有系統教授整個語文及數學課程。

二、癥結不在測驗本身

論者批評學能測驗沒有考核學生在校所學的，其實若果學校作正常語文及數學教學，全面依據固定合理的課程施教，現有的學能測驗是極爲準確，高信度及效度，且省儉的語文及數學能力指標，難找更佳替代品。

現存的問題在於部分教師或學校只針對這學能測驗施教，過度操練，其中包括極多錯誤無意義的題目，其罪不在測驗本身，而在無理操練的行爲上。舉一例子，我們可能發覺，簡單要求學生做引體向上，計算其五分鐘內的次數，已能極之準確且廉宜地比較各學校學生的體能情況。引體向上是一極佳的體能指標，但當部分學校摒棄所有體能活動，只操練引體向上，我們自然就覺得引體向上並不能充分反映學生體能。又例如，我們改變學能試內容，改爲考核作文能力，發覺教師強迫學生背誦範文，同理我們也不能因此推論作文不能反映學生的語文能力。

若仔細分析現行派位制度，不難發覺過分操練選用低劣練習，課程不均衡等問題癥結，並不在家長，《報告書》未能指出這點，反而誤導讀者，屢次將教師及家長的問題相提並論，對任何一個熟悉這派位制度的家長，若想其子女獲優先派位次序，必盡力將所有時間用於溫習校內課程及考試，鼓勵同校同學操練學能試，但自己絕不願意浪費時間於學能試題上，故對家長來說無過分操練的問題，若有，只要對其更清晰解釋派位方法，問題必迎刃而解。

小學出現過分操練，導致忽略正常教學的可能原因甚多，包括：校董會等行政干預；學校爲保證有足夠學生入讀，以求生存，故需提高學校整體表現；少數教學或行政人員欠缺專業理想；操練多項選擇題對部分教師來說，比改作文等正常教學更省

力氣。當然亦有甚多本有理想者，感到他校操練見效，在高層人士指示下，或其他原因，被迫隨俗。凡此種種原因都是源於負責操練者，並非測驗本身之罪。

非理性操練只冀求學校整體表現較佳，而不理會課程是否均衡。這困難其實可以一簡單方法解決，既然大家都認爲這些操練無助真正的語文教學，只要全港學校一起簽一個公約，協議不再做這些無謂練習，便可解決問題。學能試根本不用取消，它仍可作準確調整學校間分數之用。不過可能有部分人士仍被迫過於關注整體學校表現，違反教育原則，不參加或不遵守公約，令此計劃難以推行。

由是觀之，問題癥結在於教學決策者，家長主要關心子女的個人校內成績表現，除非受誤導，否則不會過分操練學能試題。

三、盲目追求理想測驗類型

論者亦批評學能測驗全爲選擇題，無聽、寫、理解方面的考核，並建議以語文／數學能力評估取代，其內容旨在測試學生較高層次思考方法，有別於現在的學能測驗。在語文方面，建議增加閱讀理解及作文等（這兩種方法在頗多討論會中常被引用作較佳評核方法）。

其實學能測驗內已有甚多高層次思考的題目（如：類比推理題），冀求找出數種理想題目類型

以解決問題，是不可能的夢想。以閱讀或聆聽理解題目爲例，甚至世界上最大最專業的考試中心——美國教育測量服務中心（ETS）所舉辦的大型公開考試（如：托福考試 TOEFL），也不能設計出一套完美的測驗。經驗告訴我們在 TOEFL 聆聽考試中，部分教師教授一種只看題目不用聆聽本文的方法，以達到高成績，但不能反映真正語文能力。若要考作文，不難預言，頗多小學生會被迫背誦一些範文佳句；經老師不斷訓練下，全港學生都只懂背範文而不懂真正作文，實在可悲。教育測驗學指出，若我們不斷針對及局限於某幾類型題目作練習，無論原本類型是多好及有效，最終這些題目都會失去其效用。

再者從區分學生能力來看，報告書建議新的能力評估，旨在測試學生「較高層次思考方法」，這想法在一大型公開考試中並不可行，雖然這類試題不一定全都十分艱深，但從教學的次序，我們均從熟習簡單基礎知識後，才引申至更複雜的較高層次思考，無容置疑，對某一特定課題（如：二位乘法），較高層次思考必比簡單基礎知識更艱難，故爲了解學生對某課題的能力及能準確區分各類能力的學生，我們需借助深淺各異的題目。根據傳統測量學，中等難度的題目最有區分能力，故此較艱深的高層次思考方法題目總數，客觀上不能佔太大比例。

《報告書》亦提出，考試不單局限多項選擇題，亦可包括課業類型問題。以現在一個下午考兩科的

安排，根本無可能包括課業類型問題，除非我們願意以數天時間進行這個只作校際調整用的考試。

單一課業所能考核的內容甚少，故部分學者甚至估計要多達七倍傳統測驗時間，多考幾個課業，才可比較準確評定能力，達可接納的信／效度水平。再者，因爲原有題目均爲多項選擇題，由教署內一小組以電腦處理便可，若新題目涉及甚多填空、問答、作文等類型時，所需經費不菲，報告書宜先作估計。

報告書對「多項智能」、「高層次思考方法」、「課業」（及表現性測驗 performance test）甚爲推崇。這都是有助課堂教學的較新路向，但作爲公開考試的一部分，則尙有極多問題難以解決，這正好說明公開考試的局限，只可考核一些易於測量及觀察的內容，並將嚴重扭曲小學課程。

四、「調整試導引課程改革」的問題

1. 以操練代替正常教學

如果真如報告書建議，不難預見，全港小學在五、六年級必然以大量操練模擬試題替代正常教學。甚麼「三六零」、「最新 XXX」、「最佳 XXX」、「完全 XXX」等練習均會應運而生。二十多年前的升中試惡夢極可能重現。

2. 比以前操練更多

二十年前選用學能／性向類試題，目的在於這類題目的操練效果相對較低，如今用學科題目，操

練效果更爲明顯，以前用 20%時間操練，現在則名正言順用 80%課堂時間操練。此外報告書建議公布試題，客觀效果也是鼓勵操練。

3. 放棄語文數學外的科目

外國經驗如是（Linn, 1993, 1997），香港二十多年前的情況也相同，當公開試接近時，所有其他科目均需讓路，音樂、體育、美術等非考核科目自必名存實亡，這些課堂必用於操練語文數學。就算是在語文及數學科，當公開試漸近，各國及香港經驗均指出教師不再教授與考核無關的內容（例如：英語會話），教學法也更接近考試方式，例如：辨認答案及串生字取代長篇論述題。這是我們期待的改革嗎？

4. 小學教學目標難以紙筆測試

設計一個完美測驗永遠是一個夢想，公開考試的局限更大，因爲經費、行政、測量技術等條件所限，我們只可考核那些容易觀察及計量的教學目標。在小學課程，訓練守紀律、培養求知慾、鼓勵創作力、養成正確學習態度及懂得交友相處技巧等均十分重要，「調整試導引改革」的後果必然是將種種語文數學以外的目標棄之不理。相對而言，在大學或中學後期較爲集中考核學科能力所引起的問題較少。

5. 對語文數學考核也難全面

我們也不要夢想能設計出一套理想試題去考核語文或數學，詳細理據在前節討論。簡單而言，公

開試採用的試題種類及評卷方式，所受的限制遠較校內試爲多。舉例來說，在考作文時，爲了提高改卷者的準確度（評等者信度），最容易的是採用錯字扣分法及距離要求標準字數扣分法（依標準要求字數過多或過少而扣分）等，那麼無論誰改卷，得分十分一致，若要求依創意、意境等標準改卷，則令準確度（信度）下降，熟悉公開考試的人均知道，各國要用多個評卷人去改一份作文卷，但所得分數誤差仍甚大，引來頗多爭論及上訴。日後學科試是否準備設立一龐大考試機構去管理？是否準備花費天文數字的費用去提高改卷質素？會否因種種局限導致作文（或其他同類題目）變爲背誦短文佳句？作文時教師會否只要求學生不寫錯別字，不理內容（因爲內容不影響得分）？我們是否考會話？會話（說話能力）會否被擠於課程以外？

6. 考試時間勢必加長

「調整試引導改革」若要成功，除了要解決上述種種問題外，亦必須將考試內容大量擴充，全面反映課程內容，以符合領導改革的要求。若要充分考核語文及數學，參考中學會考模式，不難想像每科 45 分鐘，一個下午考畢的「調整試」勢必改爲每科數節，數天才考完一科，這是否在報告書計劃內的改革？

7. 不應禍延他科

爲免學校單操練語文數學，一些人建議「調整試」盡可能包括其他科目，這是出於良好意願，仍是抱著以「調整試導引改革」的心態，但若我們真

的明瞭公開考試的局限，則不會贊同此舉。以健康教育爲例，老師可能需要教導學生保護牙齒、如何正確刷牙等方法，在沒有公開試壓力下，老師仍會按其專業訓練來決定教學重點，但若健教也成爲公開試考核範圍，教師必然只針對及教授公開試可以考核的部分，例如：恆齒共多少顆？公開試不可能考核學生刷牙方法，故也無需教授。「調整試引導改革」對語文數學科有負面影響，對其他科目的效果也相同。

我們同意考試有其正面效用，也有一些不良後果，但上述的問題主要源於我們希望以「調整試去導引改革」。上述分析顯示我們不單未能成功導引改革，且極可能因公開考試的種種局限，將語文及數學課程收窄於易測及可量度的範圍內，其他科目亦因大量的準備考試操練，而排拒於正常教學門外。小學課程大部分教學目標難以簡單紙筆測驗量度，以「調整試導引改革」的負面影響尤爲嚴重，其結果是與香港及世界各國的教育改革背道而馳，距離我們的理想目標愈走愈遠。

其他國家的改革及經驗

一、表現性考試

考試制度在中國有悠久的歷史（e.g., Wang, 1996），一般來說，中國父母也甚重視子女的教育及考試成績（Chen, Lee, & Stevenson, 1996; Gow, Balla, Kember, & Hau, 1996; Hau & Salili, 1990, 1991, 1996a, 1996b, 1996c; Kong & Hau, 1996; Salili

& Hau, 1994; Stevenson & Lee, 1996; Wang, 1996）。近年頗多國際研究指出，中國學童的學業表現，尤其是數學成績，均超越先進國家，這可能由於各種社會文化因素、中國父母的管教模式及重視教育所致，但亦有人認爲中國學童，只重視及集中學習一些可以具體在公開考試測量出的學習目標，故在這類考試中自然有較佳表現，但一些抽象思維、創作意念等高層次能力則極弱，不受重視（Cai, 1995, 1997），這種教學及學習態度，統稱爲應試教育，爲人詬病。

長期以來，評估方法極爲影響甚至控制教育過程，這是不爭事實。尤其是在七十年代，美國的問責潮流導致各州制定大型強制性學科測驗，以審定各州或學區的學業成就，因爲老師及學校均極爲重視這些考試的內容，加上傳媒大事宣傳考試成績之比較，不言而喻，這些學業測試被描繪成爲「課程磁鐵」（Popham, 1993a, 1993b），主宰課程改革方向。不過在八十年代，公眾及學者愈來愈關注傳統測驗對課程改革及課堂教學的負面影響，多項選擇及常模參照性的測驗最受批評，論者認爲我們需考核及重視解難、分析及較高層的思維能力（Cizek, 1993）。對於傳統測驗，學者亦關注如何制定成就標準（standard setting）、測量特性、公平性及如何編制成績報告等。

他們認爲標準測驗應作大幅修改，以符合真正教學內容及具體行爲表現，而非以常模作評核標準（Baker, O'Neil, & Linn, 1993; Bennett & Ward,

1993; Broadfoot, 1996; Hanson & Schutz, 1986; Paris & Ayres, 1994; Shepard et al., 1996）。這種評估方法可簡單統稱爲表現性、真確或另類測驗（e.g., Sadler, 1987）。它集中於仔細訂定學習目標，編制測試題目，考核及紀錄學習過程、表現、製成品及文件夾（portfolio），並決定各成就等級的指標（Madaus & Kellaghan, 1993）。學生在連續多次考核情況中的表現，被視爲較單一測試更爲重要。舉例來說：（一）學生需提交一份作業，題目爲「比較各種公共交通工具對社會的貢獻」，他們需搜集各種交通工具的數目、票價、對路面佔用情況、使用率等數據並作比較評論。（二）學生需設計並進行實驗，以比較三種紙巾的吸水能力，學者希望這類表現性測驗，能測量較高層次如解難、分析及研究能力，簡而言之，表現性測驗是著重模擬及量度被試者在真實情況下所需的能力及技巧。

表現性測驗雖然甚具吸引力，但亦有頗多局限及弱點。批評者要求表現性測驗能符合心理測量及統計要求，以便能準確將被試者分等級，並敏銳地反映能力的細微改變，它們亦應量度真正成就（Berk, 1986）。再者，在訂定合格分數或表現指標時，需合乎理論及測量效度（e.g., Broadfoot, 1995; Griffin & Griffin, 1996; Herman, Gearhart, & Baker, 1993; Huynh & Casteel, 1985; McCallum, Gipps, McAlister, & Brown, 1995; Mehrens, 1997; Messick, 1994; Popham, 1997; Schagen, 1993; Shepard, 1997; Torrance, 1993, 1995; Wolf, 1995; Zieky, 1989）。

表現性評估亦被批評欠缺公平，所測的知識不能轉移，時間及經濟上並不可行（Bracey, 1993; Linn, 1993; Worthen, 1993）。一如 Berk（1986）所總結，「釐定表現性標準常受批評，引來甚多爭論，難以執行，而差不多完全無法自辯」（p. 137）。Madaus 和 Kellaghan（1993）更爲悲觀，他們認爲「表現性測驗，漸離直接量度複雜的行爲表現，其生命不會長久」（p. 469）。

不過，爲對表現性測驗有更公平的評價，我們可能需要更廣泛的評估標準。例如，Linn, Baker 和 Dunbar（1991）相信，「嚴謹評估表現性測驗需包括：期望及非期望的結果，在特定評估所表現的能力可否轉移……，測驗的公平性……，學生解決困難時所涉及的複雜智力……，試題對學生及老師是否有意義……，判定內容質素及涵括面的基礎……及測量的成本是否合理等」（p. 20）。總的來說，雖然在方法及實施的可靠性上，表現性測驗均受批評，但這方法仍在國際上漸受歡迎及流行。

二、其他考試改革

以校本多次考核替代一次性公開考試，是世界教育改革的潮流，中國也熱烈討論及研究如何解決應試教育的問題。歐洲多國採用校內教師的多次評核作爲公開試的部分（甚至全部）分數。上海在初中及以下也採用就近入學的政策（非按成績派位）。據一項大型研究顯示（曾，1997），香港教

育的能力分隔指數是多國中最高者，甚至比新加坡還高，這顯示香港的學校能力差異甚大，能力高者集中在某些學校，能力差者則集中在另一些學校，這種能力分隔相較他國爲嚴重。

面對世界各國力求改善公開考試的種種負面影響下，香港是否抱殘守缺？甚至背道而馳？是否準備進一步加強以小學畢業公開試對課程的控制？在考慮採用「調整試導引改革」時，是否充分參考他國經驗及改革？現在用以解決學能試操練的方案，是一個進步還是倒退的政策？

從另一發展角度來看，自八零年代末開始，各國有兩個看來難以協調的教育評估趨勢（Cuban, 1995; Gipps, 1992; Gipps et al., 1995; Gipps & Murphy, 1994; Mosse & Sontheimer, 1996; Nisbet, 1993; Oakland & Hambleton, 1995; Rothman, 1995），其一是富政治性的「全國標準考核」取向，強調以全國性的基礎能力考試作監控及問責（例如：比較各校教育經費開支與學生成績的關係），其二是「表現性考試」取向，強調透過課業、工作文件夾、課題研究等形式測試，並偏重校本評估。

「全國標準考核」希望透過更多更廣泛及標準化的測驗，以提高所有學童的水平，「表現性考試」則冀求改變傳統、考核重點及方法以促進學習。前者需利用一致性高效率的測試方法，以便作跨年度、跨校、跨州、甚至跨國之比較，考試內容無可

避免地只能集中於一次性、標準化、易於執行及能準確量度的學習目標上，至於後者則希望透過多次審視學習過程中的多方面表現，重點在於全面蒐集學生各方面的能力表現，尤其是在解決接近現實世界的具體難題上，爲遷就學習能力差異，令測試成爲學習的重要環節，不同學童測試的題目及過程無需亦不能劃一，考核過程中教師亦設法依個別學生需要作不同的輔助。

由是觀之，「全國標準考核」及「表現性考試」在實施上並不一致，兩種取向有的無法協調，互爲矛盾。這問題在外國同類教育改革中早已察覺，並深明若將「全國標準考核」與生死攸關的考試結合，則難以推動「表現性考試」（Nisbet, 1993）

三、考試形式與功能的配合

國際經濟合作及發展組織（OECD）在《教育評估對課程發展》報告中，回顧多國種種考試及評鑑的經驗及影響，Nisbet（1993, p. 144）總結各國的經驗指出：

「這回顧重覆展示教育評鑑需要擔當的多種功能，因而發展出各種針對性的考核方法，不約而同的妥協方法是用不同評核方法以應付不同功能。這充滿著矛盾及衝突，如果評估依課程改革而強調改善教學，那麼我們將採用更多描述性、非審判性、不作標籤的方式去盡量支持課堂個別學童的教學，這工作主要依賴課堂教師去評鑑。不過它未能充分覆蓋整個課程，亦難以提供一些可與常模比較的分數。

相對而言，問責性的考試是一個強制性系統，要求高度可比性。易於了解及標準化的等級。這是一個外加的控制系統，與上述另一種用內部考核促進成長的方法並不相同。如果問責的考核只用於學生樣本（非整體學生），相隔一段長時間才進行一次，那麼它可以與較非正規的課堂評估相容並處……。」

測驗有各種功能，故在不同情況下所用測驗亦未盡相同。例如用標準性測驗以評核學校或整體學生水平，供問責之用；用診斷性測驗作個別學生輔導；用校內多次評核及公開試之合併成績作證書頒授；用課業習作紀錄等向家長報告（Nisbet, 1993）。

要用同一類測驗以達至學校間比較（或問責等）及教學改善的兩種功能是夢想多於實際（起碼現在如是）（見 Hau, Ip, & Cheng, 1996; Linn, 1993; Nisbet, 1993）。用於學校間比較、全國性水平等問責性考試很多時爲了減低成本，要用多項選擇式電腦可以閱卷的試題，問題一般比較表面浮淺。近年大力提倡的表現性考試，尤其是需個別或小組進行的操作考試，因費用較昂貴、難以亦不適宜合併爲單一總分、考試時間太長、評分者差異相對甚大、難用電腦批改、受時間所限難以考核課程的所有範圍（只能集中深入考核部分內容）等，故未能普遍於公開試中大量採用。

表現性試題需要投入大量專業教師的時間作爲運作成本（Impara & Plake, 1996; Saner, Klein, Bell,

& Comfort, 1994）。因爲大部分這類考試需個別考生進行，且相對考試時間甚長（可能數倍至十多倍傳統考試時間），所以若將這類考試帶進只用作調整學校間成績的機制內，爲追求更穩定的評分標準、更準確的個別學校成績、更全面的考核範圍等，在現有的測量技術下，必導致天文數字的不合理考試費用，教師用全部精力時間考核個別學生，而荒廢教學（每年九個月考核，一個月教學），本來一個推動教師多採用課業、深入探究問題的良好意願，也可能因怨聲載道而慘淡收場。

考試（或教育評核）是有多種目的及功用，故其方法及內容亦各有不同偏重。冀望公開考試能診斷個別學生的學習困難，全依賴校內評分作金字塔式的篩選，均會帶來甚多問題及困難。同理，不明辨升中「調整試」的目的，將所有教學診斷、回饋、篩選、領導課程改革等目標，強加「調整」作用之上，例如：我們本來只需相隔一段長時間（數年），隨機抽取一些學生考核，便足以作校際間調整之用，但強加其他目標，自是吃力不討好，導致大量無理操練，將小學課程嚴重扭曲。

四、以考試提高動機的謬誤

考試對課程的影響，無容置疑，但一些人甚至相信「若果沒有考試」，最好的教育制度也會失敗，不能產生果效，只要引進考試，其他問題便迎刃而解。他們相信只要有一套「正確」的獎罰制度，那怕是最懶惰、最冥頑不靈的學生也會變得努力學習。

生死攸關的公開考試不一定能提高學生學習動機（Covington & Teel, 1996; Jones, Rasmussen, & Moffitt, 1997; Kellaghan, Madaus, & Raczek, 1996; McCombs & Pope, 1994; Middleton & Goepfert, 1996; Ridley & Walther, 1995; Shade, Kelly, & Oberg, 1997; Sternberg & Spear-Swerling, 1996; Zimmerman, Bonner, & Kovach, 1996）。動機理論指出學生必需要感到考試不單是重要，而且是實質可以達致的（並非遙不可達的目標），他們才有可能增強動機。另一些學生可能認爲考試成績好壞，也無助他們日後的發展（例如：認爲人事關係更重要）。公開考試的重要性對不同年級和年紀的學生也有不同意義，對幼童來說，他們未必認同那些多年後的工作，與現在考試成績有何關係，感到壓力的只是父母（Kellaghan et al., 1996）。

公開試只會促使更多補習，注意考試技巧但並無深入了解課文內容。本來學校應該是一個學習、成長的地方，而非單去找出誰更聰明、誰較愚笨（Kellaghan et al., 1996）。在公開考試壓力下，研究顯示教師更爲專橫、強調服從、不培養學生自主、自信及創造力，公開試也令課程變得狹隘，教師及學生只關注可以用紙筆測量的教學目標，他們只集中操練那些接近過往試題的練習。

以公開考試推動課程改革及加強學生學習動機的想法，在公眾討論中屢見不鮮，不單香港部分人士有這種意圖，在其他國家同類變革中亦常出現。

美國教育研究學會爲使國民對重要的社會爭論，有最佳的專業意見及研究結論作討論基礎，特邀權威專家依這題目撰寫專書，他們的結論是（Kellaghan et al., 1996, pp. vii–viii）：

「首先，很多這類〔以公開試增強學習動機〕的建議，根本並沒有考慮動機的複雜性及公開試的可能影響，他們沒有考慮學生個人特質及背景的差異。並非所有學生均爲考試所帶來的獎賞而增加動機更爲努力。其次，那些未被考試鼓勵的，極有可能對學校有更大的疏離感。最後，就算有些學生真的因公開考試而學習，我們不難預期生死攸關的考試只會將課程收窄，只求高成績不求理解及真正掌握。這絕不能領導美國學生有更高的成就、更強的自信及創造力，或改善高階思維或解難能力。」

在公開試的壓力下，教師只依過往考試或類似試題而非正規課程去教學，學習應付考試比真正學習課程內容更爲重要，尤其在小學課程，以公開試去提高學習興趣，既不能達至目標，且遺害甚大。

可行建議及策略

一、各種解決方法

當然我們可以考慮採用一按地區更隨機派位（就近入學）的方法，但假設現行的學位分配方法不變，只考慮是否修改學能測驗，則大致上可分三種方法以改善現況：

（a） 用行政強迫或公約協議，制止學校或教師操練，得以正常教學。

（b） 以「調整試導引改革」，改變考試內容，用「理想」題目（如：《報告書》所倡議的「高層次思考」或閱讀理解），並提供樣本，公開試題，令教師操練一些有意義的內容。

（c） 讓「調整試只作調整」以種種措施減低操練的動機及報酬，盡量讓校內試主導教學。

上述方法（a）是假設教師因種種原因，未能發揮其專業決定，要用行政等方法以控制其教學。以中學會考作文爲例，私下商談，感到一般中學鼓勵學生背誦範文並不嚴重，這問題主要出現於商辦的補習學校，說明在這範疇大部分教師仍可用其專業決定，作較正常教學；相對而言，在小學操練學能考試的問題則較爲普遍及嚴重。

用行政強迫或公約協議本是最簡單的方法，但初步接觸一些行政及教學人員，認爲陽奉陰違者眾，恐怕難以實施，這方法暫時未可推行，但我們認爲最終的理想目標，仍是各行政及教學人員以其專業知識，制定適合其學生的均衡課程。

對方法（b），在前面我們已指出，問題癥結不在測驗本身，且無理想測驗類型，提供樣本試題，公開測驗也不是方法，各種衍生的流弊，在上述各節已詳加解釋，在此不贅。

我們比較贊成方法（c），因爲我們不相信有少數理想題目種類，且不受操練影響其效度。我們暫時以《報告書》的精神爲基礎，用更接近學科測驗的題目，替代現行學能試題，這可包括作文、造句、閱讀／聆聽理解、成語解釋運用、多項選擇、配對、填空等類型及內容。日後或許可推廣至會話、課業或工作文件夾等考核。

二、新建議的目標

我們的建議是基於以下的目標：

1. 盡量減低這「調整機制」對學校課程及正常教學的干擾，

2. 「調整機制」的考核內容是基於校內正常教學範圍（《報告書》的精神），

3. 「調整機制」應能反映各校學科能力的差異，

4. 盡量令花大量時間精力於操練考試技巧及某類試題的學校，感到得不償失

三、新考試內容：矩陣取樣及校本試題

「調整試」的目的在於找出某所學校在全港學校中的能力分佈情況，因該試成績並不計算在個別學生總分內，故可用「矩陣取樣」（matrix sampling）法找出學校的能力分布。簡單來說，我們設計一份甚長的試題（例如：共需十小時作答），內含多份

作文，多篇閱讀理解及一些多項選擇試題等。因每一學生作答時間不可過長（例如：一小時），故我們可隨機由每一學校抽取多隊，每隊十名學生隨機分別回答部分試題，每小隊學生的總分將可用以反映各校能力之差異，這種「矩陣取樣」法雖然涉及較複雜的統計，但在美國教育水平鑑控試（NAEP）等大型考試中採用，並證明可行。與一些 1978 年設計香港學能試的退休官員閒談，發覺當年已有這種考核模式的想法（Fischer & Molenaar, 1993; Fitzpatrick et al., 1996; Kolen & Brennan, 1995; Suen, 1990; van der Linden & Hambleton, 1996; Waltman, 1997）。

對於找出一些理想試題類別（如：高層次思維、作文、閱讀理解）用於「調整試」，從而牽引香港小學教學方法，我們未敢苟同（見上文），反之我們應盡量令學校感到「調整試」內容與校內試差別不大，無需額外操練，影響正常教學。故此我們建議先蒐集全港小五、六（甚至小四）的近年校內試題，拋掉低質素者，並作其他修訂增刪，或請經驗教師或學者再加添試題，建成一巨大試題庫，透過研究，獲取試題的難易度等測量學上特性。

所有考生用同一試卷考核是一易於操作及理解的概念，但極多國際大型考試（如：TOEFL, GRE, NAEP 等），爲了種種原因，包括跨年成績可比性，提高測量的準確度，每一考生的試卷內容及難易度毋需相同，我們可透過算式，準確地比較所有學生或學校的能力水平。在上述我們建議的矩陣取樣法

中，被選中考試的學生試卷毋需相同，一些人「作文」，一些人考「閱讀理解」及「填空」，另一些卻做「連句」及「多項選擇」，甚至聽力、說話等均可。因每人試題未盡相同，每類試題應試人數大大減少，故此一些需要較花資源（人、物力）的考試方法亦能予以採用。其實因考核範圍擴大，新方法所得的「調整指數」較以前更能準備地反映每一學校學生的能力。

四、新考試時間

學生成績主要由校內考試決定，我們需要的只是一個調整各校成績的機制，且應該不斷強調，並在真正運作上表現出這只是一個學生能力的取樣，用作調整各校試卷難易之差異，故此：

1. 學校毋需每年參加「調整試」，除了一些新校或學生水平波動甚大的學校外，基本上大部分學校可以多年才考核一次。爲免學校對某屆學生進行操練，故應以一隨機方法，不定期但平均約數年由每校抽選部分學生參加考試。在新系統運作初期，我們可容許學校選擇是否每年考試與否；選擇考試者每年要花很多精力操練，但不一定保證更爲有利，因爲成績可升可降；不考試者反而保證獲取多年平均名額，精力可放在更有意義的學習上。

2. 爲進一步減低學校能預知這調整試的時間及日期，從而進行操練，我們其實應該在年中任何時間均可抽取某校學生考核。只給學校一極短

時間的通知（如：兩星期），盡量減低對正常教學的影響，可抽取小五或小六級學生，依其在小五或小六就讀時間的長短作適當調整（例如：小五學生參加考試者將依由研究得出的方程式加分）。基本上每年十月至翌年七月均可進行考試。

因「調整試」數年才進行一次，且未知在學期哪段時間進行，既未肯定是考核小五還是小六學生，也不知會抽選中哪些學生赴考（毋需考核全部學生），故此對所有學生進行無理操練的回報太低，學校按原本自訂的教學方針進行正常教學的機會也大大提高。

五、與其他課程改革配合

報告書希望這改革能與目標爲本課程（TOC）結合，眾所周知 TOC 是要強調校內多次評核（Biggs, 1994, 1996; Education Commission Report No. 4, 1990; Lam, 1996; Lee, 1996; Report of Advisory Committee, 1994; Wong, 1996），但報告書內建議的改革是增加公開試對校內課程的影響。若要真的採用校本測驗，加強課業等表現性測驗的比重，則絕不應增大調整試的影響，只讓調整試扮演調整的角色，那麼校內試才能百花齊放，TOC 才有機會銜接。若依報告書建議，全體學生必然集中操練學科考試，那麼校內課業等評考均無望執行（Education Commission Report No. 4, 1990; Education Commission Report No. 6, 1995）。

教統會第七號報告書（Education Commission Report No. 7, 1996, 1997）亦強調優質教育的重要，並鼓勵學校有創意去提出各種校本改革方案，若操練學科試題代替正常教學，各校都拼命準備公開試，那麼誰又有興趣參與種種新的教學改革（如：大學與各小學的伙伴計劃、多讀多寫計劃）。但若果我們盡量減低調整試對學校的干擾，學校才有自由及空間參與種種教學改革。

六、應進行的研究

《報告書》提出先行研究及試辦等方法，以了解及決定是否推行語文／數學能力學科試。這看來十分科學，但並無太大實質意義。若我們明白現存問題在於學校過分關心整體成績而作操練，則除非在研究中，令學校感到有操練之必要，並觀察他們有否過分操練，否則無法研究新評估的成敗。這再次說明《報告書》迴避或未知問題的癥結。故也誤認爲這類檢驗試題（而非教學人員會否操練）的研究，可以解決當前問題。

根據本文建議，日後可做的研究包括：

（a） 翻查數據，以種種指標檢定過往多年各校在學能試表現的穩定性；

（b） 根據上述（a）的研究結果制定可行的方法，以估計多年才抽樣一次對學校派位的影響；

（c） 設計一抽樣方法及程序，以達致「調整試只作調整」用途；

（d） 設計、模擬及推算以「矩陣取樣」的具體運作，設計電腦計算程式，並以小量學校作試驗；

（e） 設計及制定各種參數，以比較多種抽樣，考核矩陣等安排對學校的影響；

（f） 設計題庫進行研究，估計題庫內各題目的測量特性（參數）；

（g） 透過研究了解小五、六年級學生在不同時段的學習情況，設計方案將不同時段（例如小五下學期及小六上學期）考核的學生，可放在同一量表上作比較；

（h） 爲確實了解整個策略的影響，選取一些學校小規模完全依據新方法作派位依據，以了解操練是否有所減少，對學校教學是否有所幫助；

（i） 制定全面推行的速度及方案。

七、新機制的公平及精確性

在現行制度，學生能否入讀其心愛學校由四個因素決定，包括：（一）個人校內試成績，（二）全校在調整試的表現，（三）在同一成績組內的運氣（即：隨機號碼），（四）所選的學校受歡迎程度。

因素（一）和（二）其實也包含運氣在內，因爲每個學生的真實能力（及潛力）永不可以準確測得。若考慮學生有多方面的能力，考試只可粗略地量度其中的一小部分，則這項目的誤差其實甚大。在因素（三），學生分爲五個人數相等組別，在同一組別，誰先選校全由運氣（隨機號碼）決定，例如：第二組別成績最好者（100 人中考第 21 名）的選校次序，有一半機會較該組最差者（第 40 名）爲後。在第（四）因素，一般來說獲取較先選校權的學生，較容易進入受歡迎的學校，但現行措施不一定百分之一百保證這選校優先次序，當某擁較先選校權的學生，因對各校受歡迎程度的資料並不充分了解，選一些與自己能力不甚匹配的學校，那麼一個同組選校次序更後的學生可能選取進入某校，而選校權較先的學生反而未能進入。故此第四因素也有涉及是否充分掌握資料及運氣因素。

由是觀之，整個派位機制雖然考慮學生能力，但整個設計根本不希望極準確地按學生能力派位，準確性從來不是考慮因素，相較 1978 年以前的升中試，現行制度是將學生能力作一定程度的混合。既然我們明瞭整個運作（尤其是第三因素）含甚大的運氣原素，那麼就算每所學校的學生能力每年有所變動，因我們採用多年才測查一次，可能導致第二因素的誤差較大，但學生在校次序仍由校內試決定，故在現行方法編入較高等級的學生，在本文所建議的新方法下，仍是被編入較高等級，若有影響（要待一些研究予以驗證），只會影響極少數在級別間邊界的學生。

由上述討論可知多年才考查一次，以數年平均作派位基準問題不大，因系統內其他部分（尤其是第二部分）刻意加進及蘊含極大運氣因素，故無必要付出高昂代價，以無理及完全干擾正常教學的操練，以換取一些根本並無需要的準確度，這就與要求以一個月時間每天考核學生，但只抽改其中一題只花五分鐘回答的題目一樣無稽。在升中派位機制中，因其他運氣因素的主宰，要求調整試有極高精確度並不合稱，以小學課程的健康及自由發展去換取這並無必要的精確度，實是本末倒置。

八、與《報告書》建議之比較

1. 「調整試」內容

我們的建議是以採用學科能力作「調整試」爲前提，但爲讓學校感到是他們主導考核內容，所以建議依據各校試卷來訂定調整試的試題，這也正面促使各校更精心設計其校內試題。考核範圍及路向也由全港學校共同制定，故提供樣題也無必要。

2. 每次考核範圍

因經費及每科考核時間所限，我們推薦用「矩陣取樣」方法，同一學校考生，將以多套試卷考核，因內容更爲全面（甚至可包含口試會話），故學校更難集中操練某幾類題目，增加操練困難。

3. 調整試的頻率

「調整試」目的在於調整，若學校成績穩定，則完全無必要每年考核，若數年才測考一次，且可由小五、六年級學生抽樣，令學校操練的對象更爲不穩定，他們付出寶貴正常教學時間去操練，但極可能並不考核該年度的學生，在操練報酬減少下，更理性地平衡操練與正常教學的機會也隨而增加。

現時因行政需要，學能試需在每年十二月進行，很多學校被迫在十二月前教完整個小六課程，連鎖效應也影響小五課程之安排，如今新考試可在小五、六任何一個時段進行，故無需全港同一時段考核，學校也無必要加緊完成課程。

4. 應試式操練

依據《報告書》建議而行，操練一定比「學能試」更爲嚴重，但若依本文方法作修改，採用「矩陣取樣」擴大考核方法，從學生群取樣，並多年才查核一次，操練情況當有改善。

5. 「調整試」之角色

本文建議「調整試」只扮演調整之角色，由校本測驗決定調整試內容及多少年才考核一次，目的旨在盡量避免這考試對正常教學的干擾，並減少操練的回報。

總 結

在具體操作上我們仍有另一建議，爲進一步減低學校對每年試題轉換，及偶然某類形式題目，對一些學校較爲陌生等所引起的焦慮，我們可取學校過往多年的平均表現作基準。每次調整試後需作改動時，除特殊情況外，保證學校新的派位在該基準的某一固定波幅之內（例如：上下限百分之十五）。簡單舉例，過往通常平均獲取 100 名一及二級名額，則不論調整試表現如何必保證有 85 至 115 名一及二級名額，每次名額變動不太大，這機制也鼓勵教師更放膽嘗試新的教學法。

《報告書》建議清楚公布確切測試範圍，提供測驗練習樣本及所有試題，使教師得知新項目種類及形式，且項目需取材自小學課程綱要等，凡此種種措施理念，都是希望以調整試導引課程改革，企圖以「理想」考試內容及種類以操縱課堂教學；但一如上文所述的各種原因，這是不切實際的夢想。與《報告書》所建議的概念剛好相反，我們不單不應公布試題以操縱教學；反之我們要主動參考各校的校內考試形式及內容，從而制定公開試的題目。我們應讓老師感到考試將全面考核所有語文／數學能力範疇，使操練者得不到相應的報酬。

部份人士希望透過調整試來改善教學，作爲問責制度的一部分，他們認爲有了這學科公開試，學校及個別教師的教學表現受考試成績所監控，對一些較不積極的教師能產生監察作用，所以公開試應每年進行。這想法可能產生甚多問題；首先用上述

多年一次的矩陣取樣方法基本上已能監控學校水平，其次我們知道公開試只可強化學生的記憶、背誦能力，一如最近訪港之美國智力研究泰斗 Robert Sternberg 教授（曾爲美國最權威心理期刊 *Psychological Bulletin* 主編）所言，如今公開試所考的能力（多爲記憶、背誦）與成人工作世界所要求分析、創造、操作等能力相距甚遠。爲了監控部分教師的工作，我們是否願意犧牲其他更爲寶貴的學習目標，浪費時間於無意義的操練之上呢？我們可否透過加諸校長或其他「質素整體視學」等系統去監察教學呢？

本文主要是依循《報告書》的一些想法，對其內容及操作作一些更具體的建議。基本上我們對考核學科作調節試內容，並不完全苟同，但我們亦深深體會學能試操練實在干擾學校正常教學，所以我們在以學科能力作調節試考核內容的前提下，提出種種策略措施，以減低這調整試對正常教學的干擾，並期望這些策略，更能反映及銜接香港及世界各地各種教育改革。

對小學過分操練的情況我們深感問題嚴重，但我們不同意設計一些理想試題，便可解決困難。我們認爲決策、行政及教學人員本身的專業知識及操守是問題的關鍵，所以就算我們對考核內容提出另一種構思，但因考試內容含更多學科成分，會否令盲目不理性的操練更爲厲害？會否令社、科、健、勞、音、體等科目讓路，以提供更多操練時間？最終我們是否要以同區某種抽籤方法（就近入學），

才可令教學納入正軌？這些都令我們深以爲慮。我們對教師專業及自主性仍抱有熱切期望，並認爲這是對小學課程及教學較理想的發展方向，以此信念，我們提出上述各項批評及建議。我們認爲整個計劃需以漸進方式推行，並不斷監察各校實施情況，以免由火坑掉進另一深淵。

參考文獻

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*, 1210–1218.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*, 137–172.

Biggs, J. (1994). What are effective schools? Lessons from East and West. *Australian Educational Researcher*, *21*(1), 19–39.

Biggs, J. (Ed.). (1996). *Testing: To educate or to select? Education in Hong Kong at the crossroad*. Hong Kong: Hong Kong Educational Publishing.

Board of Education. (1997a). *Report on review of 9-year compulsory education*. Hong Kong: Sub-committee on review of school education, The Board of Education, Hong Kong Government.

Board of Education. (1997b). *Report on review of 9-year compulsory education (revised version)*. Hong Kong: Sub-committee on review of school education, The Board of Education, Hong Kong Government.

Bracey, G. W. (1993). Assessing the new assessments. *Principal*, *72*(3), 34–36.

Broadfoot, P. (1995). Performance assessment in perspective: International trends and current English experience. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment* (pp. 9–43). Buckingham, UK: Open University Press.

Broadfoot, P. M. (1996). *Education, assessment and society*. Buckingham, UK: Open University Press.

Cai, J. (1995). *A cognitive analysis of U.S. and Chinese students' mathematical performance on tasks involving computation, simple problem solving, and complex problem solving*. Reston, VA: The National Council of Teachers of Mathematics.

Cai, J. (1997). Beyond computation and correctness: Contributions of open-ended tasks in examining U.S. and Chinese students' mathematical performance. *Educational Measurement: Issues and Practice*, *16*(1), 5–11.

Chen, C. S., Lee, S. Y., & Stevenson, H. W. (1996). Academic achievement and motivation of Chinese students: A cross-national perspective. In S. Lau (Ed.), *Growing up the Chinese way: Chinese child and adolescent development* (pp. 69–92). Hong Kong: The Chinese University Press.

Cizek, G. J. (1993). Reconsidering standard and criteria. *Journal of Educational Measurement*, *30*, 93–106.

Covington, M. V., & Teel, K. M. (1996). *Overcoming student failure: Changing motives and incentives for learning*. Washington, DC: American Psychological Association.

Cuban, L. (1995). A national curriculum and tests: Consequences for schools. In *The hidden consequences of a national curriculum* (pp. 47–62) (A public service monograph). Washington, DC: American Educational Research Association.

Education Commission Report No. 4. (1990, November). The Curriculum and behavioural problems in schools. Hong Kong: Hong Kong Education Commission.

Education Commission Report No. 6. (1995, December). Enhancing Language Proficiency: A comprehensive strategy (Part 1 & 2) (Consultation document). Hong Kong: Hong Kong Government.

Education Commission Report No. 7. (1996, November). Quality School Education (Consultation document). Hong Kong: Hong Kong Government.

Education Commission Report No. 7. (1997, September). Quality School Education. Hong Kong: Hong Kong Government.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1993). *Rasch models: Foundation, recent developments, and applications*. New York: Springer-Verlag.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, *33*, 291–314.

Gipps, C. (Ed.). (1992). *Developing assessment for the national curriculum*. London: Kogan Page.

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995). *Intuition or evidence? Teachers and national assessment of seven-year-olds*. Buckingham, UK: Open University Press.

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham. UK: Open University Press.

Gow, L., Balla, J., Kember, D., & Hau, K. T. (1996). Learning approaches of Chinese people: A function of socialization processes and the context of learning? In M. Bond (Ed.), *The handbook of Chinese psychology* (pp. 109–123). Hong Kong: Oxford University Press.

Griffin, M. M., & Griffin, B. W. (1996). Situated cognition and cognitive style: Effects on students' learning as measured by conventional tests and

performance. *The Journal of Experimental Education*, *64*, 293–308.

Hanson, R. A., & Schutz, R. E. (1986). A comparison of methods for measuring achievement in basic skills program evaluation. *Educational Evaluation and Policy Analysis*, *8*, 101–113.

Hau, K. T., Ip, M. H., & Cheng, Z. J. (1996). Target oriented curriculum and inter-school comparison. *Education Journal*, *24*(2), 1–13.

Hau, K. T., & Salili, F. (1990). Examination result attributions, expectancy, and achievement goals of Chinese students in Hong Kong. *Educational Studies*, *14*, 17–31.

Hau, K. T., & Salili, F. (1991). Structure and semantic differential placement of specific causes: Academic causal attributions by Chinese students in Hong Kong. *International Journal of Psychology*, *26*, 175–193.

Hau, K. T., & Salili, F. (1996a). Achievement goals and causal attributions of Chinese students. In S. Lau (Ed.), *Growing up the Chinese way: Chinese child and adolescent development* (pp. 121–146). Hong Kong: The Chinese University Press.

Hau, K. T., & Salili, F. (1996b). Prediction of academic performance among Chinese students: Effort can compensate for lack of ability. *Organizational Behavior and Human Decision Processes*, *65*, 83–94.

Hau, K. T., & Salili, F. (1996c). Motivational effects of teachers' ability versus effort feedback on Chinese students' learning. *Social Psychology of Education: An International Journal*, *1*, 69–85.

Herman, J. L., Gearhart, M., & Baker, E. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, *1*, 201–224.

Huynh, H.; & Casteel, J. (1985). A comparison of the minimax and Rasch approach to set simultaneous passing scores for subtests. *Journal of Educational Statistics*, *10*, 334–344.

Impara, J. C., & Plake, B. (1996). Professional development in student assessment for educational administrators: An instructional framework. *Educational Measurement: Issues and Practice*, *15*(2), 14–19.

Joint Committee on Review of Secondary School Place Allocation Mechanism. (1996, December). *Report on the review of secondary school place allocation mechanism* (in Chinese). Hong Kong.

Jones, B. F., Rasmussen, C. M., & Moffitt, M. C. (1997). *Real-life problem solving: A collaborative approach to interdisciplinary learning*. Washington, DC: American Psychological Association.

Kellaghan, T., Madaus, G. F., & Raczek, A. (1996). *The use of external examinations to improve student motivation* (A public service monograph). Washington, DC: American Educational Research Association.

Kolen, M. J., & Brennan, R. L. (1995). *Testing equating: Methods and practices*. New York: Springer-Verlag.

Kong, C. K., & Hau, K. T. (1996). Students' achievement goals and approaches to learning: Relationship between emphasis of self-improvement and thorough understanding. *Research in Education*, *55*, 74–85.

Lam, C. C. (1996). *Target Oriented Curriculum: A dream which will never come true?* (in Chinese; Occasional Paper No. 1). Hong Kong: Hong Kong Institute of Educational Research, The Chinese University of Hong Kong.

Lee, P. T. (1996). *Response to TOC assessment guideline*. Hong Kong: C.C.C. Primary School Principals Association.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, *15*, 1–16.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, *16*(2), 14–16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researchers*, *20*(8), 15–21.

Madaus, G. F., & Kellaghan, T. (1993). The British experience with 'authentic' testing. *Phi Delta Kappan*, *74*, 462–469.

McCallum, B., Gipps, C., McAlister, S., & Brown, M. (1995). National curriculum assessment: Emerging models of teacher assessment in the classroom. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment* (pp. 57–87). Buckingham, UK: Open University Press.

McCombs, B. L., & Pope, J. E. (1994). *Motivating hard to reach students*. Washington, DC: American Psychological Association.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16*(2), 16–18.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Middleton, J. A., & Goepfert, P. (1996). *Inventive strategies for teaching mathematics: Implementing standards for reform*. Washington, DC: American Psychological Association.

Mosse, R., & Sontheimer, L. E. (1996). *Performance monitoring indicators handbook* (World Bank Technical Paper No. 334). Washington, DC: The International Bank for Reconstruction and Development.

Nisbet, J. (Ed.). (1993). *Curriculum reform: Assessment in question* (OECD Documents). Paris, France: Centre for Educational Research and Innovation, Organisation for Economic Co-operation and Development.

Oakland, T., & Hambleton, R. K. (1995). *International perspectives on academic assessment*. Boston, MA: Kluwer Academic.

Paris, S. G., & Ayres, L. R. (1994). *Becoming reflective students and teachers: With portfolios and authentic assessment*. Washington, DC: American Psychological Association.

Popham, W. J. (1993a). Circumventing the high costs of authentic assessment. *Phi Delta Kappan*, *74*, 470–473.

Popham, W. J. (1993b). Educational testing in America: What's right, what is wrong? A criterion-referenced perspective. *Educational Measurement: Issues and Practice*, *12*(1), 11–14.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13

Report of Advisory Committee on Implementation of Target Oriented Curriculum. (1994). Hong Kong: Education Department.

Ridley, D. S., & Walther, B. (1995). *Creating responsible learners: The role of a positive classroom environment*. Washington, DC: American Psychological Association.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Francisco, CA: Jossey-Bass.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*, 191–209.

Salili, F., & Hau, K. T. (1994). The effects of teachers' evaluative feedback on Chinese students' perception of ability: A cross-cultural and situational analysis. *Educational Studies*, *20*, 223–236.

Saner, H., Klein, S., Bell, R., & Comfort, K. B. (1994). The utility of multiple raters and tasks in science performance assessments. *Educational Assessment*, *2*, 257–272.

Schagen, I. P. (1993). Problems in measuring the reliability of National Curriculum Assessment in England and Wales. *Educational Studies*, *19*, 41–54.

Shade, B. J., Kelly, C., & Oberg, M. (1997). *Creating culturally responsible classrooms*. Washington, DC: American Psychological Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–8,13,24.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice*, *15*(3), 7–18.

Sternberg, R. J., & Spear-Swerling, L. (1996). *Teaching for thinking*. Washington, DC: American Psychological Association.

Stevenson, H. W., & Lee, S. Y. (1996). The academic achievement of Chinese students. In M. Bond (Ed.), *The handbook of Chinese psychology* (pp. 124–142). Hong Kong: Oxford University Press.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum.

Torrance, H. (1993). Combining measurement-driven instruction with authentic assessment: Some initial observations of National Assessment in England and Wales. *Educational Evaluation and Policy Analysis*, *15*, 81–90.

Torrance, H. (1995). Teacher involvement in new approaches to assessment. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment* (pp. 44–56). Buckingham, UK: Open University Press.

van der Linden, & Hambleton, R. K. (Eds.). (1996). *Handbook of modern item response theory*. New York: Springer-Verlag.

Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement*, *34*, 101–121.

Wang, G. (1996). Educational assessment in China. *Assessment in Education*, *3*, 75–88.

Wolf, A. (1995). Authentic assessments in a competitive sector: Institutional prerequisites and

cautionary tales. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment* (pp. 88–104). Buckingham, UK: Open University Press.

Wong, C. K. (1996). *Comments on TOC assessment guideline*. Hong Kong: S.K.H. Primary Schools Council School Principals' Association.

Worthen, B. R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan*, *74*, 444–454.

Zieky, M. (1989). Methods of setting standards of performance on criterion referenced tests. *Studies in Educational Evaluation*, *15*, 335–338.

Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). *Developing self-regulated learners: Beyond achievement to self-efficacy*. Washington, DC: American Psychological Association.

To Avoid Jumping from the Fire Pan into Fire: Recommendations on Assessing Primary Students' Academic Ability

HAU Kit-tai

(Abstract)

In a recent review of the 9-year compulsory education, the Board of Education recommended the adoption of an Academic Ability Assessment (AAA) to replace the present Academic Aptitude Test (AAT) in the allocation of primary students to secondary schools. We agree that the coaching on the AAT using low quality exercise is undesirable but the idea of using AAA to guide the primary school curriculum is also detrimental. It is argued that there is no ideal question type and that the AAT may lead to intensive coaching on a very narrow and limited syllabus, which cannot reflect the great varieties of more important educational goals and objectives. On the assumption that an academic achievement based test will replace the existing aptitude test, we review relevant educational reforms in Hong Kong and other countries and propose some useful strategies that we should consider. This includes the implementation of the matrix sampling design in which students within a school are tested with different sets of questions, so that a much wider range of content and question types (e.g., oral examination) can be used. Furthermore, we believe that the moderation examination (i.e., the AAT or AAA) should limit its function to moderation and should have a minimum interference on school teaching. Thus, for schools with stable quality of student output, which are the majority, we should also consider assessing these schools once every few years. These strategies are statistically valid and in line with school based examination, performance type assessment, Target Oriented Curriculum and other educational reforms.