目　　錄

表　目　錄

圖目錄

壹、前言

一、參加源起

人才為國家發展之本，先進國家在提升國家競爭力的過程中，特別重視人才的養成、人才的選拔、人才的在職訓練以及人才是否符合職場的專業需求，此即考選部所重視的教、考、訓、用合一的政策重點。本部依法辦理各項公務人員暨專門職業技術人員之考試，其主要任務在甄選各類專業人才與常任文官進入各類專門領域與政府部門，為一般民眾提供服務。欲有效達成前項目標，則需善用有效評量工具與評量方法，以選出優秀適任之人才。我國的國家考試大多採取筆試作為評量工具，大多偏向「分科成就測驗」，但以應試科目作為主要評量效標的作法，仍有可討論之處，尤其依據應試科目所命擬之試題，是否可有效區辨應考人的核心職能，而核心職能是否又能完全符合用人機關的需求，均值得進一步探討。測驗工具有選擇決定與安置決定的功能，爰此，不斷提升測驗工具的信度與效度以及採用多元評量方法應是考選部的施政主軸。他山之石，可以攻錯，先進國家發展科學化測驗工具的經驗，考量我國的文化適應性後，應有助改進我國人才考選與精進評量方法。

國際測驗委員會（International Test Commision）係由各國心理學會、測驗委員會及相關測驗機構所組成的非營利組織，主要功能在於推廣及評估測驗政策之制定、研發與實施。該會成立於1978年，總部設在英國，目前有20個正式會員（國家及專業心理協會）與46個副會員（研究機構、測驗出版商與其他測驗委員會）和超過 150 位個人會員（學生及從事測驗工作之個人）。約在 10多年前，開始在世界各國輪流舉辦年會，每兩年辦理一次，提供國際性的平台，讓各國學者發表其研究成果，也提供發展測驗之公司介紹研發之測驗工具的機會。

本部為強化國際參與及吸取先進國家發展測驗工具之經驗，

於 2007 年 10 月以副會員身分加入該會，並自 2008 年第 6 屆年會開始派員參加，藉以了解國際測驗發展趨勢，促進測驗科學及其應用之交流與合作。2010 年國際測驗委員會年會於香港中文大學舉行，除有歐、美國國家測驗學者參與外，亞洲各國如新加坡、日本、香港、中國大陸、印尼、菲律賓均有學者前往發表論文，國內學者與研究生參與者也不在少數。會中的論文發表以及各種公開場合或私下之討論，對我國國家考試評量方法、效度之提升與電腦化測驗的發展，也激盪出一些想法與改進措施。本報告謹就參加本次年會所見所聞，選擇部分與我國國家考試有關之重要主題，除說明其主要內容外，並提出心得與建議，以作為未來改進國家考試測驗工具與評量方法之參考。

二、行程紀要

會議行程表

日期	時間、交通資訊	地點	會議及活動
7/17 （六）	10:20 長榮航空 BR867 12:00	台北至 香港	啟程 下午赴香港商務書局購置考試試題彙編
7/18 （日）	9:00-12:30 13:30-17:00	香港中文大學	參加測驗工作坊
7/19 （一）	9:00-12:30 13:30-17:00	香港中文大學	1.報到暨開幕式 2.各場次研討會
7/20 （二）	9:00-12:30 13:30-17:00	香港中文大學	各場次專題研討會
7/21 （三）	9:00-13:30 19:20 長榮航空 BR870 21:00	香港 至 台北	1.各場次專題研討會 2.閉幕

2010 年國際測驗委員會年會因在香港舉行，在經費允許的情況下，由本部試題研究中心黃科長立賢與資訊管理處余分析師慶杉共同前往參與，除於會議召開前參加「測驗工作坊」外，因大會發表之研究論文超過200篇，專題演講也超過30場，僅能選擇與本部業務有關之議題參加。

三、2010年國際測驗委員會年會紀要

2010年第七屆國際測驗委員會年會於7月18日至21日於香港中文大學舉行，這是年會首度離開歐洲到亞洲首度舉辦，計有超過40個國家及400位學者、測驗公司人員和政府機構代表參加，活動內容分為兩部分，一為8場次的「測驗工作坊」，本部人員選擇參加之場次為「不同試題功能與偏誤-探討其概念、方法與應用」、「國際測驗委員會的測驗品管準則-計分、試題分析與測驗分數的報告」；二為專題研討會，進行的形式包括專題演講、論文發表與研討、海報展示與詢答。本次年會的主題為「全球經濟架構下測驗和評量發展的挑戰與機會」（Challenges and Opportunities in Testing and Assessment in a Globalized Economy），會議共有五項議題如下：

（一）國際性測驗的心理量與測驗理論發展。

（二）本土、第二語言和跨國的測驗發展。

（三）利用科技精進測驗實施與資料管理。

（四）跨國測驗的發展的政策、倫理、專業技術和訓練。

（五）跨國測驗發展的安全性和隱密性。

在短短三天會議中，探討的主題相當廣泛，無論是專題演講或是論文研討會，有許多新的研究成果值得我們注意，尤其是各種測驗科學化的研究進展與測驗「效度」的提升。唯因國情不同，國外編製測驗通常有大型測驗公司的介入，測驗學者則負責開發與評估使用成效，我國則由政府機關辦理，僅有少部分引進測驗學者進行研究。爰此，在赴港參與會議之前，先依大會所提供之參考資料，選擇與本部辦理業務有關之場次，若有疑問處則與該場次主講人討論或索取詳細之資料。本研究報告分為國際測驗委員會「測驗品管準則」、多元評量方法的發展、測驗的安全與資料的隱私、電腦適性化測之議題與應用，說明其研究內容與發展趨勢，並提出相關結論與建議。

貳、國際測驗委員會「測驗品管準則」

國際測驗委員會自1994年開始，為了控制測驗品質的一致性，即努力於建立一套跨文化、語言可以適用的「測驗品管準則」(***Guidelines for Test Use***)，目的在確保測驗計分、分析和分數報告的品質控制（ITC，2000)。發展的主要原因，第一在於有許多國家已有專業性的心理計量組織與人員，有些國家則否，必須藉由一個完整的機制，在測驗的編製與使用上控制其品質；其次，目前紙本與網路普遍使用的測驗工具經常違反測驗的使用倫理與隱密性；第三，許多不同的國家為因應勞動力需求增加，發展職業性向測驗領域，希望在不同執業領域選去具有發展潛力的工作者；第四，近年來的網路使用改變了職業測驗的施測與計分方式，同時影響管理測驗的標準化流程與隱密性。經過10多年的發展，測驗品管準則的建立已經較為完整（Allalouf，2010)，以下分別就一、準則內容介紹、二、一般性的準則、三、工作準則的步驟加以說明之：

一、準則內容介紹

（一）目標

在測驗發展的開始、管理、行政上的應用、計分、測驗分析、測驗解釋與報告，「正確性」是最基本的要素。從錯誤答案所導致的計分錯誤，原始分數轉化為標準分數的錯誤、藉由電腦計分所得到的測驗分數錯誤、突發性的對於應考人分數紀錄錯誤，此類情形都不應該發生。在編製測驗或使用過程中，在不同階段有關的人或組織，都有責任保持測驗的標準化與專業性，例如用人單位、心理計量組織、學術機構、政府機關。測驗的專業使用者應該能夠了解及預測測驗編製與運用的每一階段中，可能發生的錯誤樣態，除了能偵測錯誤外，更能避免錯誤的發生。測驗的專業使用人士若能具備測驗品質控制的廣博

知識，將能確保測驗的正確性。

（二）測驗品管準則運用的對象

測驗品管準則目的在於確保測專業性與正確性，尤其在測驗發展的每個階段都應依準則進行各項工作，各階段包括：

1. 測驗的管理。
2. 測驗的計分。
3. 編製試題和試題分析（包括試題標準化和等化）。
4. 測驗解釋。
5. 分數結果的報告以及提供給測驗採用者或編製者。
6. 對測驗使用者進行訓練和督導。
7. 設計電腦化資料管理系統處理測驗資料。
8. 制定測驗使用的政策（包括立法者熟悉測驗的運用規則）。
9. 測驗的出版。

測驗品管準則對於上列的事項著墨甚多，不但吸收了許多測驗實務上的經驗，同時也參照許多測驗評量實證研究的成果。

（三）測驗品管準則的必要性與測驗常犯的錯誤

無論是心理能力測驗、學習成就測驗或是職業性向測驗，在測驗與評量的計分、分析與報告領域中發生錯誤，都是非常嚴重的問題。譬如說，假如有極大比例的計分錯誤，除降低測驗信度外，預測效度也同時偏低。在某些實例中發現，人格測驗計分錯誤會導致測驗受試者被認為偏離常態，具有變態行為；職業性向或學科能力考試的錯誤計分，常會造成具有具有能力的應考人落榜，或導致入學考試的錯誤安置。錯誤的計分也常常會誤導各種教育計畫的介入，或導致專門職業技術人員證照考試錯誤的發照。總之，測驗中發生各種錯誤會導致嚴重的後果，可能會導致公共利益的損害，以及一般應考人對舉辦考試的信賴度，或者導致法律上的爭訟，以及錯誤用人以後仍需重新聘用具有能力的工作者，造成用人機關的損失。

（四）測驗品管準則的範圍

測驗品管準則涵蓋大規模測驗使用機構與多種試題型式(例如選擇題、實作評量、結構性和非結構性的口試、評估團體活動的表現)。另外在使用目的上也相當多元，例如對於教育上的目的或是職業性向能力的評估。測驗品管準則對於「個人化的測驗」或「團體化測驗」都有相當明確的規定，同時有廣泛的應用性，無論是能力測驗、成就測驗、性向測驗、臨床上的診斷測驗都具有實用性。

測驗品管準則適用於任何形式的測驗管理，例如紙筆測驗、電腦測驗（網路測驗或電腦單機測驗）或其他模式的測驗。而有效的品管準則應用在測驗計分、分析與報告中，並需有需同時考量測驗的建構效度、信度與結果的預測度上。分配各項資源在測驗品質的管是一種負責任與提升公平性的做法，也是各項工作領域重要的倫理規範。

（五）來自於其他專業領域的實例

品質管控流程常運用在其他專業領域中,每一種品管流程都有其獨特性,學習其他專業領域的經驗有助於測驗評量領域品管流程的建立，例如工程、飛行、電腦軟體和醫藥領域。與醫學有關的領域，常常會注意到發生在醫院的錯誤，將這些錯誤修正，成為有貢獻的方式。例如藥品的保存過程不適當、醫療處置的不恰當、新式醫療器材的操作失誤、有瑕疵的溝通、不良的團隊合作和缺乏清楚的安全政策。前述原則同理可證，測驗錯誤的產生，可轉換為考試試題的存取過程發生錯誤，或是測驗分析與解釋的錯誤。

（六）測驗品管準則的架構

1. 一般性的準則：優先考量測驗的計分、試題分析與分數的報告。
2. 品管準則的各項步驟：每一項步驟的指南（step-by step instruction)。

二、一般性的準則

（一）確認目前正在使用的品管標準

1. 若有必要的話，決定目前在測驗中使用品管準則，形成測驗品管的必要性程序，同時回顧、更新和監控目前的準則是否合宜，而且不時地作例行性的檢查。
2. 在管理測驗時，確認在測驗的每一個步驟確認其適當的標準。
3. 當處理一個新測驗時，必須執行一個詳細的計分、分析與報告的查核過程，同時確認現有的標準涵蓋所有過程。
4. 定期更新與修正品管的準則。
5. 在發展一個新式測驗時，建立明確的具體的測驗品管準則。

（二）對於所有參與測驗過程所有的人基本的準備與共識

1. 測驗發展與應用過程中，每一個環節所參與的人都應對所負責任取得共識。
2. 決定與陳述各種測驗的使用目的，例如篩選的目的、測驗成就的目的或是研究的目的。
3. 建立測驗使用者或團隊最佳的溝通方式，例如傳達相關訊息從一個團隊到另一個團隊；或是傳達詳細的描述從測驗發展團隊到測驗分析團隊。
4. 關於測驗的過程，應與受試者建立良好的溝通管道。
5. 決定或測量資料的方法，例如紙筆測驗藉由光學讀卡機或掃描器獲得作答資料，電腦化測驗用透過電子化的管道讀取資料。
6. 測驗具有決定的功能，但是需確認各種分測驗（subtests）的權重與分數所代表的意義。
7. 對於測驗計分規則取得共識，也就是對正確或錯誤答案在數量上都有精確的計算。

8. 選擇計分表和決定量表分數的範圍。

9. 決定如何處理未回答的資料是相當困難的，例如應考人有些試題未作答、測驗評估者忘記對特殊考生作特別評量，因為重新進行評量是不可能的。

10. 若進行不同測驗的等化與分數比較，則需界定和描述等化的設計和方法。

11. 對於應考人與測驗單位而言，分數所代表的意義、分數的分布應該呈現至何種程度，是值得討論的議題。

12. 在目前法律的規定下，決定應考人測驗的結果，那些測驗機構應該承擔後續的處理與保管工作。

（三）測驗的資源

1. 確認有足夠的資源能夠應付測驗計分、分析與報告所需。

2. 檢查每一種資源的備份，例如測驗專家不能作等化工作，誰可以替代？或是測驗題答案卡光學讀卡機故障，如何解決讀卡之問題？

3. 假如測驗的運用者在未預期的情況下缺席，緊急的備選方案應立刻啟動。

4. 決定時間資源的需要性，建立測驗計分、分析與報告的時間表建立每一個步驟所需的確切期限是可行的。

5. 決定電腦和網路資源的需要性，例如個人電腦數、維持的系統、硬碟的容量、伺服器的空間、網路頻寬等。

（四）利害相關人的需求和期望

應考人注意應試的過程是否符合他們的需求和期望，尤其是分數等化的過程以及取得分數報告所需的時間。前項需求和期望，介於測驗利害相關人事之間，應該是合理與交互溝通的。有關進一步利害相關人的需求和期望說明如下：

1. 利害相關人之間，如測驗編製銷售者、測驗使用者（應考人）以及測驗結果應用者之間形成一個適當的契約關係。

2. 當發生問題時，對於解決方案達成共識，例如選擇題產生答案的爭議或是應考人在作答過程中遭到吵雜環境的影響。

3. 對於單一選擇題，事先原本設定只有一個正確答案，經過考試後卻產生超過一個以上的正確答案，都需要建立標準的處理流程或是避免此類問題的發生頻率。

4. 在已經完成計分程序後，卻發生試題的錯誤，需建立此類錯誤的處理程序。

5. 提供應考人對於所公布之試題與答案挑戰更正的機會。

（六）專業的工作人員和工作氣氛

假使測驗的編製與應用過程對於專業的團隊是相當重要的，對於新雇用的測驗專業人員除了其專業性考量，維持組織氣氛和諧也相當重要。

1. 確認測驗計分、等化和報告的專業人員具備不可或缺的知能。

2. 免給予個人工作表現的速度的額外壓力。

3. 避免過多的額外加班。

4. 嘗試在測驗的編製、使用和報告的過程中，創造穩定、放鬆，又能避免產生錯誤的組織氣氛。

（七）證明及報告錯誤

1. 發現錯誤的方法，可以在測驗進行過程中的每一個階段，使用標準化的檢查表，確認在每一個階段產生錯誤時應負責的工作成員。

2. 發現每一個階段錯誤發生的原因與本質，並採取有效的處理測策略。

三、工作準則的步驟

（一）測驗結果的報告設計

對於測驗機構或是應考人只給予測驗分數是不夠的，適當的測驗分數解釋是非常重要的。測驗的發展、計分和分析階段都應將最後的結果產出納入考量，在測驗發展之初，就要確認分數的報告能夠使應考人充分了解其意義。因此，不同面向的分數解釋在一開始就要被提出，除了單一分數外，其餘從原始分數所衍生的其他量化數據所代表的意義都應解釋清楚。

（二）資料背景

背景資料在品質控制上的應用有幾項目的，確認應考人的特質、解釋未預期的結果、為了等化的目的建立適配的比較團體，以下介紹各步驟：

1. 假如法定的程序允許，蒐集各項背景資料，例如性別、種族、教育程度、過去測驗的分數。
2. 在報名參加考試之前，就應透過個人填寫或電子化的方式蒐集個人背景資料。
3. 系統性且周期性的檢查應考人歷次報名及作答的歷史性資料，同時注意到應考人重複接受測驗後作答資料的不一致性。
4. 在背景資料與分數之間進行研究以了解其相關，例如應考人的年齡是否與取得測驗高分有關。假如某一測驗證明年紀較大之應考表現顯然會高於年較輕的應考人，當測驗進行發生相反的結果，即年紀輕的應考人成績優於年紀大的應考人，計分過程則需檢現是否發生錯誤。

（三）計分

所有應考人辛苦的作答應該以電子化的方式存檔，而且每一位應考人的答案都有足以辨認的編碼，紙本或電子化存檔的應考人作答資料均應有法定的保存期限，以應付專業與法律爭訟的需

要。

1. 紙筆測驗的作答資料需留存數年，年限長短則依照每個國家的法律規定。
2. 關於電子化的儲存，需使用不斷電系統與備用電池或其他科技的方法，而且這些科技的技術不能造成資料的損害。
3. 使用掃描器時，應該定期的調整準確度，同時建立操作手冊。
4. 建立嚴格確認應考人身分的系統，例如身分證號碼重複出現，或通過測驗後仍重複報考。
5. 所有的資料應該被安全的保存，同時資料可分開保存，例如一種是歷史性的應考人背景資料，一種是與應考人身分對應的作答資料。

測驗資料在資料庫中應妥善的運用及儲存，應考人的答案通常以原始分數的方數計算，「古典測驗理論」通常只以正確或錯誤的答案作為計算原始分數的依據，但猜測率未被估計，部分試題的權重通常也未特別計算。「試題反應理論」關心的是潛在能力，利用θ值或潛在分數表示應考人得分狀況，應考人得分情形常會因為許多不同類型的錯誤受到影響，尤其一些錯誤導致偏低的得分，使用下列的方法可避免錯誤：

1. 以描述性統計比較測驗常模的分數與樣本分數的範圍，若對應考人樣本進行統計分析時，需察覺應考人能力的內在變異，同時大規模的樣本應特別注意潛在能力的估計。
2. 再度審視極端分數（極高分與極低分），有可能在獲得資料之初就產生錯誤。
3. 檢視其他不適合的指標。
4. 若測驗當中有子科目或是分測驗，需進一步了解應考人再這些科目或分測驗中的差異。
5. 檢查試題的統計特徵，從試題當中去發現試題可能有的疑義或是試題鑑別度的高低。

6. 對於不同測驗狀況下的應考人給予不同的關注與核對資料正確性。

（四）測驗分析

1.測驗題與開放性測驗的評分分析

在作為行政決定時，試題分析可提供有關試題特徵的統計資訊，整體分數的組合結果往往可呈現應考人完整的表現。除非應考人的樣本過少，否則每一種測驗的試題分析資料具有參考價值。古典測驗理論的試題分析資料包含難度與鑑別度，依據反應理論進行的試題品質分析則使用參數估計來了解測驗使用的模式。除此之外，試題分析可提供信度與標準誤、平均數與標準差、應考人的成績常態分布和作答反應。當測驗應考人超過最低人數後，下列程序應該特別注意：

（1）運用具有可信度的電腦試題分析系統。

（2）當考試辦理後進行試題品質的分析作業。

（3）對應考人的成績作結論時需先審視試題品質分析資料。

（4）進行試題品質分析前，需先行確認答案的正確與否，若有新的答案均需予以更正。

2. 對於實作測驗、工作樣本、角色扮演與口試進行評分

鑒於選擇題型的試題被認為客觀與具有可信度，開放性題型的試題，例如實作評量、開放性的問卷、工作樣本、角色扮演經常都有主觀的成分在內，後者的主觀性高於前者，因為開放性題型的試題比選擇題的信度低，包含太多的人惟評分及個人經驗因素，但也有許多方法可以使用，降低試題計分原有的主觀性，以及改善計分的信度與正確性。

（1）確認實作測驗、工作樣本的辨認、角色扮演與口試都能藉由具有專業知識和經驗以及經過訓練取得證書的評分者負責。

（2）對於評估開放性試題的評分指南應該是清楚且具有結構

性。

（3）在進行評分過程前，評分者應接受訓練，使評分者熟悉評分的各項作業程序。

（4）開放性的測驗評分，評分者至少應為兩名。

（5）若使用電腦進行開放性測驗評分時，仍有有一位評分人員責監控整體的評分過程。

（6）評分過程中，評分者的評分應是獨立不受干擾的。

（7）應用統計程序評估信度和評分過程，計算評分者評分的一致性以及彼此間的差異。

（8）假如評分者不符測驗舉辦機構的期望，例如評分信度過低，或與其他評分者的差異過大，除在評分過程中面告外，未來也無須猶豫的替換其他評分人員。

（9）發展處理評分不一致的政策，當評分者分數的差異小時，分數可以被平均；若是差異過大時，有經驗的評分者應該調整其差異。

3. 等化新的測驗與試題

假如分數需要比較，不同測驗間的各項心理計量特徵就需要比較，等化又分成前等化和後等化過程，可使用試題難易度、量尺方式進行比較。另外可使用試題反應理論中不同的方法進行等化。

（1）如果有非預期性的等化問題發生，例如分數偏低，需先確認在相同的標準化狀況下，所有測驗的類型與試題都能夠被處理。

（2）發展並詳細確認等化的標準程序。

（3）探討等化程序的基本假設，並確定不同的基本假設是否導致相類似的結果。

（4）比較不同應考人背景的分數，假如有任何的不一致性，需確認原始分數。

（5）若要訂定應考人的及格或不及格分數，可將前幾年的分數應考人的背景和相類似的測驗納入考量。

4.計算標準分數

在許多實例中，標準化後的分數使人更容易了解與應用，例如標準九和其他標準分數。參數和轉換後的量表被使用來計算標準分數與百分位數，成為成績單中的主要內容。

（1）將原始分數作適當的轉換，獲得清晰的的分數表。

（2）確認低原始分數與高原始分數轉換成為標準分數的正確度。

（3）在某些情況下，界定標準分數的最小值與最大值有其必要性。

（4）比較新測驗與其他測驗的量表與參數，尤其是差異性與相似性。

（5）計算隨著時間過去，在測驗量表分數間的改變。

（6）比較人工計算成績與電腦計算成績的差異。

（7）檢查原始分數與標準分數的相關可使用剖面圖呈現。

5.測驗安全性的檢核

如應考人測驗的分數被發現是使用不誠實的方法得到，例如作弊，將導致嚴重的後果。但作弊不能在測驗前完全被發現與禁止，特別在可獲得高利益測驗中，作弊的誘惑力是巨大的。在與應考人作弊的對抗中，可以諮詢律師以及試務工作人員，確保試題的安全性與處罰的適用性。以下是一些先前的提醒與建議：

（1）考試進行前座位的安排，特別注意不要將熟識的應考人的座位安排在附近，可以應考人名字的字母順序隨機安排。

（2）雇用可信賴的監考人與例行性的訓練。

（3）檢查與常理不合或未符預期試題反應型態，例如簡單的

試題正確率過低，困難的試題回答正確率也過高。

（4）確認應考人身份，必須使用照片或生理特徵的確認方法（例如指紋或瞳孔的掃描）。

（5）在考試之前獲得應考人手寫的字跡樣本，幫助分辨代考者。

（6）分析重複參加考試者的分數差異，藉由統計分析了解差異的合理性解釋。

（7）處理作弊的應考人需要使用法律的手段，同時也需事前制定影響考試公正性的各種偷竊試題或作弊行為處理的政策。

（8）使用專屬的隱藏性廚櫃保存測驗的資料和結果。

（9）使用電腦進行測驗資料的儲存和傳送都需受到限制。

（五）測驗結果的報告

測驗分數公布給應考人和測驗使用者得知時，理想上，分數應有紙本與電子成績單，網路目前已經成為普遍且標準的成績報告方法。應考人和測驗使用者必須了解成績單的呈現方式。

1.任何拿到成績單的應考人都能對呈現的分數有適當的了解。

2.產生電腦化的分數報告有助於拿到成績單的應考人了解測驗的實際表現。

3. 清楚了解不同分數所產生的意義，也需注意某些子分數（部分科目分數）可信度不足作出高風險的決定。

4. 使用應考人「焦點團體」、「有聲思考法」、「實驗研究」與「一對一的訪談」，以協助測驗機構對於分數解釋的內容與解釋應如何設計。

5. 成績單切忌塗改，如有印刷錯誤就應重新印製。

6. 對於成績單的儲存和轉換分數都應該在電腦上設定密碼。

參、多元評量方法的發展

本次年會除探討成就測驗的評量外，其他心理測驗的論文發表也不在少數，尤其是「人格測驗」與「情緒智力測驗」。非成就測驗的評量方法在入學考試或就業考試較少使用，以人格測驗而言，人格的測量涉及人格定義、信度與效度、反應心向與偽裝答案的問題，所使用的字句對不同人有不同的意義，語意上的差異造成人格測量的困難。至於 EQ(Emotional Intelligence，稱為情緒智力)，從廣義觀之，其意謂個人自我掌握以及人與人之間圓融互動的能力或人格特質，其涵蓋範圍，譬如如何激勵自己愈挫愈勇（自我驅策力）；如何克制衝動與遲延滿足（自制力）；如何調適情緒，避免因過度沮喪影響工作能力（熱忱）；如何設身處地為人著想（同理心）。以下謹將本次年會對人格測驗與情緒智力測驗的重要研究與發表結果整理如下：

一、人格測驗

（一）如何應用人格測量的五大人格特質

Costa & McCrae（1989)所提出的分類法，在 1980年代 Costa＆ McCrae在馬里蘭國際健康組織中對老人現象進行研究，然後研究發現五種具代表性的人格因素（BIG FIVE）定名為：1.神經質；2.外向性.3.開放性；4.親和性；5.勤勉正直性，各構面的定義和特徵的說明如下：

1.親和性：代表容易相處、溝通且與人合作，舉止行為相當有禮貌。

2.勤勉正直：代表自我努力、專心和集中程度。又包含自我要求、追求卓越、循規蹈矩、謹慎、有責任感等特質。

3.外向性：代表自信、主動活潑、喜歡表現、喜好參與熱鬧的場合、喜好結交朋友、活潑外向等特質。

4. 神經質：代表激起一個人負面情感刺激的強度，得分高者代表容易緊張、憂鬱、沮喪、情緒化，缺乏安全感。

5. 經驗開放性：代表一個人對於事實及新奇事務的吸收與好奇程度，得分高的人代表好奇、富有想像力、喜歡求新求變的特質。

五大人格特質概念，經常被提出運用在成人人格的了解上，但同時也常有四項爭議，第一，人格特質的測量，「五項」是否足夠？到目前為止實證性的資料是否足以支持其他特質的形成？第二，不同人格特質在行為的反應差異性，測驗所使用的試題，是否足以測出所代表的人格特質？第三，對於預測組織人員的工作表現，人格特質測驗相當有用，但對於區辨效度與倫理等議題仍需加以驗證與關注；第四，人格特質測驗跨文化的議題相當重要，尤其測驗使用在全球化的浪潮下，不同文化下的測驗題目應有所不同與修正。

（二）五大人格特質測驗與何倫碼測驗的相關與變異

特質因素論的假設認為個人特質與工作要求條件假如相互適配，可找出理想工作生涯。因此又稱適配論。以一組特質或人格特質來界定不同類型的人，同時也以一組工作上所要求的條件或資格來界定不同類型的工作。此處所謂的特質指個人的人格特性，包括：性向，興趣，成就，價值觀和個性等，可經由測驗或量表等工具加以測得。

John Holland 的生涯類型論，認為生涯選擇係個人人格在工作世界中的表露和延伸；人們係在工作選擇和經驗中表達自己，個人興趣和價值。個人對自我的觀點，與其職業偏好間的一致性，即構成 Holland 所稱的典型個人風格（金樹人，2009），以圖形表示可成為何倫碼六角形。何倫類型論典型（Holland Code）個人風格與典型職業的個人人格個性與適合職業說明如下（江文慈，2005）：

1. 實際型（R）：此類型的人具有順從、坦率、謙虛、自然、堅毅、實際、有理、害羞、穩健、節儉等特徵。其行為表現為：（1）喜愛實際操作性質的職業或情境；（2）以具備實用的能力解決工作或其他方面的問題；（3）擁有機械和操作的能力，較缺乏人際關係方面的能力；（4）重視具體的事物或明確的特性。典型的適合職業為工程師、工程人員、醫師、醫；事技術人員、農、林、漁、牧相關職業、機械操作員、一般技術人員。

2. 研究型（I）：此類型的人具有分析、謹慎、判斷、好奇、獨立、內向、精確、理性、保守、好學、有自信等特徵。其行為表現為：（1）喜愛研究性質的職業和情境；（2）以研究方面的能力解決工作及其他方面的問題；（3）擁有數學和科學方面的能力，但較缺乏領導才能；（4）重視科學價值。適合的職業為數學家、科學家、自然科學研究人員工程師、工程研究人員、資訊研究人員。

3.藝術型（A）：此類型的人具有複雜、想像、衝動、獨立、直覺、創意、理想化、情緒化、感情豐富、不重秩序、不符權威、不重實際等特徵。其行為表現為：1.喜愛藝術性質的職業或情境；2.以藝術方面的能力解決工作或其他方面的問題；3.富有表達能力、創造能力、擁有藝術、音樂、表演、寫作等方面的能力；4.重視審美價值與美感經驗。適合的職業為音樂家、畫家、詩人、作家、舞蹈家、戲劇演員、導演、藝術教師、美術設計人員。

4. 社會型（S）：此類型的人具有合作、善意、慷慨、助人、仁慈、負責、善溝通、善解人意、富洞察力、理想主義等特徵。其行為表現為：（1）喜愛社會性質的職業或情境；（2）以社交方面的能力解決工作或其

他方面的問題；（3）具有幫助別人、瞭解別人、教導別人的能力，但較缺乏機械與科學能力；（4）重視社會規範與倫理價值。適合的職業為一般教師、神職人員、輔導諮商人員、社工人員、護理人員、社會服務工作者。

5. 企業型（E）：此類型的人具有冒險、野心、抱負、樂觀、自信、有衝勁、追求享樂、精力充沛、善於社交、說服他人、獲取注意、管理組織等特徵。其行為表現為：（1）喜愛企業性質的職業或情境；（2.）以企業方面的能力解決工作或其他方面的問題；（3）具有語言溝通、說服、社交、管理、組織、領導方面的能力，較缺乏科學能力；（4）重視政治與經濟上的成就。適合的職業為業務行銷人員、企業經理、公關人員、政治人員、律師、法官、媒體傳播人員、仲介代理人員。

6. 傳統型（C）：此類型的人具有順從、謹慎、保守、自抑、謙遜、規律、堅毅、實際、穩重、重秩序、有效率等特徵。其行為表現為：（1）喜愛傳統性質的職業或情境；（2）以傳統方面的能力解決工作或其他方面的問題；（3）具有文書作業與數字計算方面的能力；重視商業和經濟價值。適合的職業為會計師、會計人員、總務、出納、銀行行員、行政助理編輯、資訊處理人員。

目前許多職業探索測驗均依據何倫碼與五大人格特質架構發展其測驗，許多大規模的跨文化研究調查存在於不同研究組別中的「不變性」與「差異性」。Neal（2010）的研究發現，分別以何倫碼與五大人格特質測驗作為研究工具了解其人格內在組型，發現其差異性甚小，對於預測即行為表現也有極佳的功能，兩項測驗工具研究結果的相關程度非常高。

（三）**五大人格特質測驗的選才功能**Fruypt（2010）的研究發現過去十五年來，在工商心理學、人事心理學與組織心理學領域中，使用五大人格特質的理論架構所發展出來的人格測驗成為一股潮流。在人格測驗發展出的軌道中，也藉助近年來人格測驗的研究成果，例如「潛在特質活化理論」、「個人中心途徑」、「特質概念化光譜」，實務上則蒐集各項職業領域傑出表現者的資料，並研究其人格特徵為何。

（四）五大人格特質測驗文化上的差異

荷蘭的幾位學者Serlie、Hiemstra、Van Leeuwen與Bazen等人在本次年會中所發表的論文，旨在討論五大人格特質文化的差異。在美國與歐洲，工作表現與人格特質已討論甚多，以問卷為形式的測驗結果在選才過程中扮演重要的角色。若以不同語言進行施測，受測團體中可能有少數人員不了解甚至誤解題目的意義，可能產生各種偏誤，測驗公平性的問題也經常被提起。為了探究測驗偏誤與公平性的問題，測驗學界近年來使用不同試題測驗功能統計比較的方式，矯正此種誤差（Different Item Functioning ，簡稱DIF)。目前的研究已著手使用不同DIF的方法檢驗五大人格特質測議的文化偏誤，我們的假設是因種族的文化差異會導致反映結果的差異。以荷蘭280位大四或碩士在學最後一年的學生作為受試者，以五大人格特質架構所發展的FFM人格測驗作為問卷，受試者分為兩組，一組母語為英文者；一組母語為荷蘭語，研究結果發現在神經質的人格特質偏誤上約有5.5% 的比例，其他人格特質的測量偏誤並未達顯著標準。

二、情緒智力測驗

「智力」是多年來心理學家關心的議題之一，從最早1905年的「比西量表」發表至今，已超過百年的歷史。心理學家也隨著時代的變遷與演進，研發不同評量內涵的「智力測驗」。然而，最先將「情緒」與智力併在一起討論，提出「情緒智力」概念的是Salovey與Mayer（1990），他們認為情緒是由社會智力的概念發展而來，與Gardner在多元智力理論所提及的「人際智力」及「自知智力」較為類似。所謂「人際智力」是指能夠認知他人的情緒、性情、動機與慾望，並能夠做適度的反應進而與人交往且和睦相處的能力；「自知智力」則是自我認知的基礎，係指能夠認識自己的感覺，認識自己並選擇自己生活方向的能力。情緒智力的概念提出後，引起眾多心理學家的興趣，在Goleman出版「EQ」一書後，對於情緒智力的定義、內涵與測驗編製紛紛出現各種相關的學術研究與討論。

國內外學者對於情緒智力測驗的內涵與分類眾說紛紜，以下將先就二十年來國內外重要的情緒智力測驗量表或問卷架構的文獻加以整理，再將本次國際測驗委員會的研究發表論文加以說明。

（一）歷年重要的國內外情緒智力測驗量表評量內容

1. Goleman（1995）的「情緒智力測驗量表」

   本量表用來測量情緒能力、社會能力與人格特質等，包含認識他人情緒、自我激勵與處理人際關係等分量表。

2. Copper 和 Sawaf（1997）的「情緒智力地圖自我評分版本」

   本量表包括「情緒智力地圖問卷」、「情緒智力地圖評分方格」及「情緒智力地圖詮釋指引」等三部分。情緒智力地圖用以了解與情緒智力相關的各種成分和能力，包含「目前的生活環境」（生活事件、工作壓力與個人壓力覺察）、「情緒辨別率」（情緒自覺、情緒表達、對他人情緒的覺察）、「情緒能力」（意圖、

創造力、韌性、人際關係、具建設性的不滿)、「價值觀和信念」(同情心、人生觀、直覺、信賴範圍、個人力量、誠實)、「情緒智力成效」(健康狀況、生活品質、人際商數、最佳表現)等五個構面。

3. Simmons(1997)的「情緒智力簡要自我測量表」

本量表內容包含十三種主要性格方面的情緒智力類型，分別為情緒精力、壓力處理、樂觀性、自尊、工作承諾、注意細節、改變的慾望、勇氣、自我導向、果斷、容忍、考慮他人、社會傾向。

4. Weisinger(1998)的「情緒智力量表」

本工具分為三部分，第一部分在協助受試者評定情緒智力的能力，內容包括自我能力(自我覺察、管理情緒、與自我激勵)和人際能力(與人相處良好、情緒教導)；第二部分是受試者自己檢查在問卷上的回答，了解本身的優點和需要改善的地方；第三部分是利用四周時間練習與觀察本身情緒智力方面的技能，再重新作第一部分和第二部分的練習，註記出與之前差異之處，發現自己進步的情形。

5. Mayer、Caruso和Salovery(1999)的「多因素情緒智力量表」

本問卷包含十二個量表，由「情緒界定」、「情緒使用」、「情緒了解」、及「情緒管理」等四大構念組成，情緒界定係指用來知覺與辨識各種不同刺激的情緒內涵，計有情緒臉譜測驗、心情音樂測驗、圖片設計測驗、情境故事測驗；情緒使用係指情緒同化與情緒使用的能力，計有感覺調和測驗與感覺偏誤測驗；情緒了解係指理解情緒表達的能力，計有複雜情緒測驗與相對情緒測驗；情緒管理係指管理情緒的能力，計有他人情緒管理測驗與自己情緒管理測驗。

6. 王春展(199)的「兒童情緒思維智力量表」

參酌相關情緒智力理論編製，包含「自我情緒智力」與「個人

情緒智力」兩層面，每一層面又分為情緒察覺、了解、推理、判斷、調節、激勵與反省等七個成分。

7.江文慈（1999）的「青少年情緒量表」

本量表共分成四個層面，包含（1）「情緒察覺」：能分辨內心的各種情緒、覺察真實感受的原因、察覺別人的情緒、了解別人內心的感受；（2）「情緒表達」：適切地表達出自己內心的情緒感受，面對別人的情緒亦能適切因應；（3）「情緒調整」：克制情緒衝動，特別是在面對人際衝突與憤怒挫折時，能夠使用一些調整策略改善情緒狀態的強度，以紓解不舒服的情緒或維持正向的情緒；（4）「情緒運用」：運用情緒訊息來思考、選擇、計劃或解決問題，從經驗中獲得啟示，提升自我成長。

8.王財印（2000）的「國中學生情緒測驗智力量表」

本量表依據 Mayer 與 Salovery（1997）所修正的情緒智力概念編製而成，旨在測量國中學生的情緒智力，包含（1）「覺察、評估及標答情緒」；（2）「激發、產生及促進情緒」；（3）「了解、分析及運用情緒知識」；（4）「反省、調整及提升情緒知識」等四個分量表。

9.王佳玲（2004）的「大學生情緒智力量表」

依據 Mayer 與 Salovery（1997）所提的情緒智力構念，將情緒智力分為「情緒界定」、「情緒使用」、「情緒了解」與「情緒管理」等四種成分。

（二）香港國際測驗年會情緒智力測驗重要的發展與研究

自從情緒智力的概念提出後，在實務界與科學界吸引許多的注意，雖然在實務界相當盛行，但在科學界卻受到許多批評。使用自陳量表尋求情緒智力特質的途徑，常常缺乏區辨效度；使用最大表現能力測驗探求情緒智力能力的途徑，在計分與試題內容上遭到批評，以下有四個有關情緒智力最新的研究方法與發現，分別說明之：

1.評估中國人情緒智力反應的方法

中國學者Wang,Chi-Sum等人(2010)使用四種方法評估情緒智力，共計進行十二年，他們認為情緒表達應被視為一種能力而非人格的表現。第一種方法是藉由效標關聯效度得到與情緒智力證據；第二種方法是使用其他受試者熟悉且經過測試的試題，在法律的保護下，藉由問卷施測得到信度與效標關聯效度的證據；第三種方法是利用受試者對於預試試題正確與不正確的反應，建構情緒智力的正式試題；第四種方法是利用口試或面談蒐集信度與效度的證據。

2.使用多元方法評量情緒智力的研究與發展

有兩種評量情緒智力的方法經常使用，第一種即所謂的「情境判斷測驗」（situational judgment test ），利用編製完成的劇本評量個人遇到特殊情境的反應；第二種方法是「主要代理人」範例的測驗，從安排的情境與情緒反應中，將受試者比擬為故事中的主角，在碰到模擬情境時，會產生何種情緒與行為，並請受試者預測其後來的發展。在一個受試人數達857人的研究中，受試者為社區大學與一般大學學生，評估其心理特質，並了解不同團體的變異（例如種族不同），共進行情緒智力測驗（使用多元情緒智力測驗）、壓力測驗及五大人格特質測驗，同時也了解其再測信度，研究結果發現多元情緒智力評量工具是相當可信的。

3.評量情緒知識建構計分的關鍵

情緒智力最主要的議題是有關「能力」的題目如何「計分」，也就是如何能夠對於情緒反應的題目確認是正確優良的反應且給予高分。有三種方法常用來計分，即「一致型的評分」、「專家的評分」與「目標性的評分」，但前述評分方法也常遭到挑戰與批評，很難有效作為辨認情緒智力題目有效或反應正確的方法。目前另有一種計分方式，利用「情緒知識」的探究了解情緒智力的構念，六個構念主成分中，共有142個特質，六個構念分別為

評價、表達、主觀經驗、身體反應、動作傾向、調整程度，利用情緒的表達字彙與每日的情緒事件及情緒特質加以對應。此種方法曾經應用在一個北京、瑞典和英國的跨國研究中，利用情緒字彙的意義可以預測四個構面，包括個人愉悅性、潛能、自我激勵與其他不可歸類的能力（Fontaine & Scherer，2010）。

4.評量情緒智力牽涉的概念

我們在日常生活中發現，教師對學生、愛人之間、父母對孩子、朋友之間的行為，常藉由判斷對方情緒反應以進行各項社會互動。對於每日週遭的人自動化的使用語言和非語言的行為表達我們的情緒，同時人們也常常不用花費太多的心力，就能夠對他人表現的行為予以解碼。目前的研究方式也有採以「情緒定錨」與「情緒調適」作為情緒智力的方法，藉由判斷情緒表達的內容，作為判斷情緒溝通腳本的依據。

肆、測驗的安全與資料的隱私

本次年會重點之一為「進行全球性測驗時所應考量的安全與隱私問題」，該議題主要是因為許多企業或組織均比以往更加重視具有技能、資質及知能的國際人才發掘，眾多認證計畫，如資訊技術、醫療或金融業等，舉辦範圍屬於全球性，而大學與學院的招生計畫也接受來自世界各地的學生申請，企業職前篩選考試也希望能招聘到具有全球水平潛力的員工。正因大量全球流動性的勞動人口與學生人數，以及利用網際網路從事市場經銷、通訊、教育及評鑑等重要因素，促使測驗的範圍逐步朝向國際性發展的趨勢。

由於這些具有重要決定或高風險性的評量（high-stake test），影響著受測者重大的權益，因此如何有效的管理測驗時的安全，與保護受測者的隱私，成為本次會議的討論重點。

一、測驗之安全省思

在重要決定性的評量（如檢定、證照、職前篩選、徵才、入學等）中，測驗的品質相當重要，因它左右整體考試結果的公平與公正性。一般足以影響測驗品質的威脅有下列幾種：

（一）偷竊或欺騙者使用新式的技術，如從傳統的抄襲、手機、微型對講機、針孔攝影機、無線耳機，到能傳送答案的手錶、眼鏡、隱形筆等。

（二）欺瞞老師、家長、雇主等道德淪喪的比例逐年升高。

（三）考取與否的利害關係加重。

（四）過時的安全模型。

（五）含糊不清的評量計畫。

根據美國新聞與世界報導調查指出：

（一）百分之八十的「高成就」測驗中，高中學生承認作弊。

（二）百分之五十一的高中生不認為作弊是錯誤的。

（三）百分之九十五的作弊高中學生說，他們沒有被發現作弊。

（四）百分之七十五的大學生承認作弊，百分之九十的大學生不相信作弊者會被抓到。

（五）百分之八十五的大學生表示，作弊一定要成功。

顯然作弊在測驗中是經常發生，然而影響測驗品質的威脅雖相當眾多，但歸納受測者的行為後發現，受測者的行為可區分為作弊（Cheating）與剽竊（Piracy）兩大類。如何有效防治呢？會議中有篇研究（Foster，2010）提出相當另類的觀點，也許可從賭場（Casino）的作法得到借鏡。因為在賭局中，總是爾虞我詐，欺騙與偷竊行為層出不窮、手法日新月異，而且這些行為是持續進行、永不間斷，故它所提供的安全防護是最具領先地位。

在說明賭場中的最新安全防護措施前，先介紹「十大通緝作弊者」名單，藉由每位應試小偷名字之給定，以突顯常見之作弊流行手法：

(一)冒用者（The Impersonator）：假冒他人身分，嘗試誤導監考人之代考者。

(二)走私者（The Smuggler）：違法攜帶測驗物品、材料或設備的受測者，這些設備包括隱藏的相機、計算器、手機、小紙條、水壺，以及將答案寫在鞋子、手、腿等部位。

(三)說書人（The Storyteller）：藉由測驗之參加以背誦試題並轉述給他人之受測者。

(四)連鎖幫（The Chain Gang）：串連多位說書人，有策略、系統的透過網際網路等方式收集、銷售試題內容之集團。

(五)時間旅行者（The Time Traveler）：在全球測驗環境中，有許多考試都在同一天進行，因此利用跨時區特性，非法共享試題或提供答案供較晚測驗之時區者，即稱為時間旅行者。

(六)合作者（The Collaborators）：聘請程度優良之合作者透過訊息傳遞手法，以提升自己的測驗成績者。

(七)羅賓漢（Robin Hood）：監試者、老師或測驗管理員基於某些原因，可能修改弱勢人員或貧困學生之答案者，以提升其成績排名。

(八)駭客（The Hacker）：對於未受到適當的保護措施或未進行內部一致性檢查的測驗系統，在未經授權的情況下，滲透該系統，以偷取測驗試題或答案等內容，甚至修改考試最終的成績。

(九)賣黃牛票者（The Ticket Scalper）：企圖在免費的預試測驗中回答不誠實的答案，以誤導試題的難易度、鑑別度等因子或測驗試題的內容，或收購試題內容予以轉賣者。

(十)內神通外鬼者（The Insider & The Fence）：測驗機構內部人員或消息靈通之人員事前偷取測驗試題內容，並交由販售試題的人予以銷售試題給相關受測者。

接下來則介紹應用在賭場中的最新安全防護措施，以及它為測驗所帶來的省思：

(一)車牌號碼辨識器：在賭場中所提供的車牌號碼辨識器具有自動化、特定區域及會員資料庫連結等特性，若應用在測驗中，可以採取的類似作為，包含生物辨識（臉孔、聲音、指紋、虹膜等）、電腦/櫃檯式的身分驗證、測驗的歷史紀錄連結等預先防禦作業。

(二)錄影設備：在賭場中會進行全程的錄影，以達到對賭客的嚇阻作用，並作為事後稽查的依據，因此在測驗的環境中，我們可思考設置現場的監考官或自動偵測工具（如網路攝影機），透過平移與縮放等動作，即時監看受測者的施測過程。

(三)無線射頻識別系統（Radio Frequency Identification, RFID）技術：RFID 應用早已存在你我日常生活中，像是捷運悠遊卡、高速公路 ETC 等，它具有嵌入式、隱避的特性，故在

測驗上的應用，可以採取數位式的試題浮水印或木馬試題等技術，以避免網際網路自動化搜尋功能，並可快速定位被盜取試題，有效保護試題的內容。

(四)通信介面：賭場中會用它與客人進行溝通或同事間的合作與集中控管，而應用在測驗上，則監試者可與受測者進行必要的交談（如操作說明或事件排除）、存取受測者過往的測驗紀錄、提供受測者即時且客製化的考試等。

(五)繳費窗口：這是賭場的最後一道防線，以作為分析與決策的最終參考，若在測驗上的應用，則應加強資料的採證、事件的審查及測驗分數的驗證等方面。

二、證照考試應有之安全考量

由於證照考試通常關係著受測者的重大權益，例如可否執業、要求加薪、轉換較佳職場等，因此在考試過程中，較常會面臨的安全挑戰包含考試舞弊具有相當大的商機、難有效採取法律救濟、科技發展所衍生之問題、擴大考試實施範圍造成執行上之困難及作弊者之間互相學習等。

為確保考試舉辦之公平性，Fremer（2010）提出有效防止考試作弊或試題竊取等舞弊事件，應事先制定防弊相關策略，如建立測驗之安全計畫、嚴加控管所有與安全相關之事務（如政策、程序及工具）、使用網路監控、分析測驗資料等多重方法，並瞭解與明訂相關權責，如課程管理與考試服務交付之廠商、操作面與法律面、相關單位間（如檢調單位）之溝通、領導者的需求、高階主管的支持等。

一般要增強組織資料保衛之安全性，必須明確描述資料管理政策與受測者間之聯繫，該政策至少應包含下列內容：

(一)受測者條約應包含法規的限制，例如測驗結果應保存多久、受測者必須在何種期限內提出成績複查申請等。

(二)測驗管理者應妥善儲存受測者結果並保留一段指定的時間，以便事件發生時可立即收集、調閱相關資訊，例如人口統計資訊、調查結果、考試管理報告等。

(三)成績應該在確認受測者均符合資格且合格之後才得以報導。

(四)如果其中一項或一群測驗結果的完整性在無法被確認前，應先扣留測驗成績與相關決策，直到已確認完成該資訊之準確性。

(五)有關測驗結果與受測者相關資訊之資料共享原則，應確保訊息的完整性與考量受測者應有之權益。

另外當增建與記錄資料之安全時，應考量下列管理程序：

(一) 在測驗管理前、中、後，有關測驗成績與測驗成績的處理應確實遵照政策規定辦理，以確保其完整性。例如測驗管理者如何取得日誌（log）與事件等報告。

(二)程序中應明訂答案卷、受測者之履歷及受測者身分識別等相關文件之安全傳輸與處理方式。

(三)增加掃描答案卡等非電子形式之測驗結果處理程序，以降低潛在的欺詐行為，如塗銷或修改答案等足以影響實際作答結果之事實。

(四)所有電子資訊均應使用加密與安全傳輸等技術。

(五)測驗結果的存取應被適當的管控與監控，亦即程序中應說明未經授權的存取與篡改資料等行為。

(六)應定期審查測驗程序，以發現安全漏洞與其可能對測驗、成績及決策方案所造成的影響。

(七)所有試題、測驗之產出文件、測驗結果及受測者個人資訊等資料，其歸檔作業應遵循業務需求與安全政策之一致性方式辦理。

(八)處理與執行程序應依據測驗結果的處置需要制定，尤應確保個人隱私與資料之保護。

一些可採取之較具效率整合式方法，包含：

(一)安全的稽核，用以檢視是否尚有測驗安全上之漏洞，甚至發現可再加強之地方。

(二)網路之監控，用以確保試題是否有暴露的情形，或是否有違反相關安全規定之行為。

(三)資料採證，用以瞭解試題暴露的數量，以決定是否進行題庫試題之更新與管理。

一旦發生安全事件時，就應當機立斷，立即採取處理行動，以防止事件擴大，下面幾起國際間近期在測驗過程中發生較重大安全事件之真實案例、處理措施可作為借鏡。

(一)2010年7月12日FSBPT宣布暫停特定國家之所有畢業生的NPTE考試。

(二)2010年6月9日美國醫學內科部決定醫生可從審查過程中發展新試題之措施。

(三)2008年6月23日GMAT網站上因提供學生錯誤的指引而遭罰300萬美元罰款。

三、就業測驗應有之安全考量

本次會議中有篇報告（Burke, 2010）介紹了一項在全球 50 多個國家中進行的人才招募線上測驗系統，該系統是由英國 SHL 所發展，該公司成立於 1977 年，主要是提供雇主僱用具有良好性格與能力之評估服務，特別的是於 2004 年即開始施行線上就業測驗，共提供 26 種語言服務，超過 500 萬人次使用。

目前該系統所驗證之範圍包含數值、口語、歸納、檢查、計算等能力，希望藉由一連串的標準測驗，提供雇主較適合人選，而該項線上就業測驗較傳統式人才招募模式，具有下列優點：

(一)節省時間與費用：因早期招聘過程較欠缺效率，使用線上測驗可較快速篩選出適合人選，且更優質的聘用結果，可降低員工流動率，提高員工的整體表現。

(二)更好之聘用決策：因該錄用之決定與否係依據相關可靠性之數據，以消除與減少舞弊、身份驗證及違反安全規定等風險，且所採取之先進且嚴峻的驗證過程，更可準確地預測出受測者真正的能力。

(三)易於使用：在考試管理中，所需介入之行政資源相當低，且可輕易整合各式招聘方案；另開放每天 24 小時之測驗服務，受測者可以選擇在最方便的時候與地點，隨時透過網路來進行測驗。

因該評量採線上測驗的形式，故其最大特色是沒有測驗中心的概念，而且無人在現場監督的情形下進行考試，如此安全將是一大威脅與挑戰。因此為有效確保考試過程中的安全，SHL 發展出測驗的安全框架（見圖 4-1、安全框架）與驗證程序（見圖 4-2、驗證程序），也就是透過網路巡邏的方式，查看是否有涉嫌違反安全規定的行為。若發現有違規情形，則應採取證據蒐集法制單位、與嫌疑犯聯繫、資料採證等必要處理程序，最後評估有那些

題庫試題遭受洩漏情形，以便進行試題內容的更換。

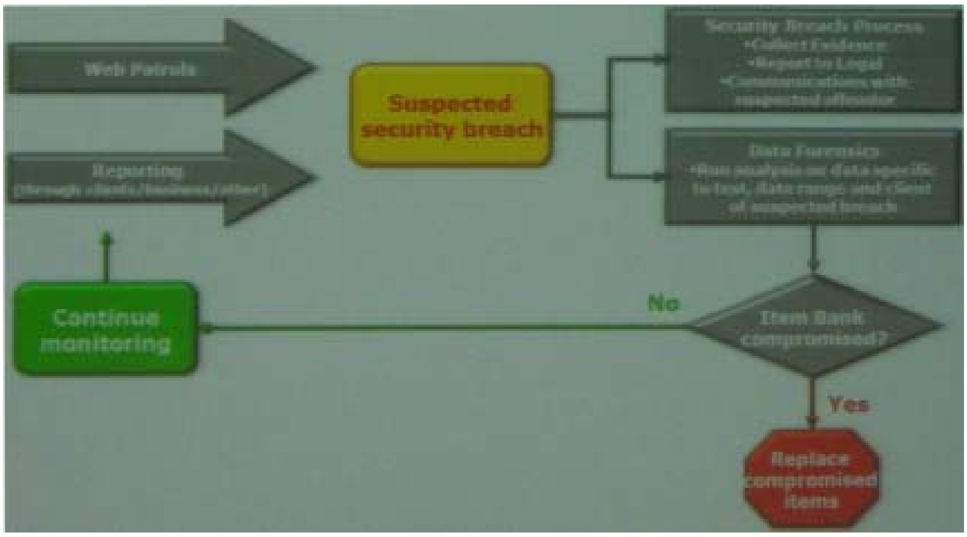

Web Patrols
Reporting
(through clients/business/other)
Suspected security breach
Security Breach Process
•Collect Evidence
•Report to Legal
•Communications with suspected offender
Data Forensics
•Run analysis on data specific to test, date range and client of suspected breach
Continue monitoring
No
Item Bank compromised?
Yes
Replace compromised items

圖4-1、安全框架

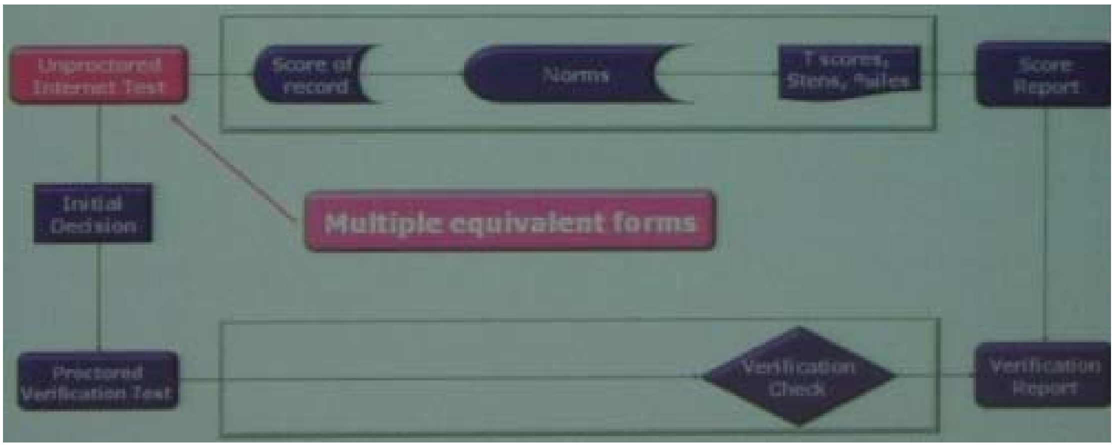


圖4-2、驗證程序

總之，測驗的安全不只是技術層面的問題，更是心理層面的問題，因為我們必須讓受測者與相關利害團體相信，整個測驗過程具有積極管理的程序且安全無虞。

伍、電腦化測驗之應用

本次會議中有許多議題均與電腦化測驗相關，包含電腦化適性測驗，近年來國際間電腦化測驗之發展趨勢，發展電腦化測驗可參考之準則與標準，以及日本分享該國近幾年在電腦化測驗上之應用，相關議題說明如下：

一、電腦化適性測驗

電腦化適性測驗係以試題反應理論（Item Response Theory, IRT）為學理基礎，於 1952 由 Lord 首倡此一理論，其後學者再加以發揚光大。IRT 主要是描述題目參數、受測者能力與其作答反應機率的數學模式，基於 IRT 的單向性（unidimensionality）與局部獨立性（local independency）假定，只要試題符合 IRT 的模式，則接受不同難度試題的受測者其能力是可以互相比較的。

近年來由於電腦效能大幅提升與網際網路的蓬勃發展，國際間已愈來愈多採用試題反應理論作為編製測驗、施測、計分、解釋及提供諮詢服務的依據，成功跨出電腦化測驗發展的一大步。其迷人之處主要是電腦化適性測驗具有傳統紙筆測驗所具有的功能、特性、優點或測驗結果，且測驗時間遠比紙筆測驗節省更多的施測時間。

然而電腦化適性測驗仍存在著一些問題有待解決，例如如何加強試題校準的準確性、如何控制試題的曝光與試題重複率、如何偵測出受測者的欺騙行為等，本次會議中，有一篇研究（Wu, 2010）即探討是否可能使用小樣本，即可有效且精準的估算出試題參數。

試題反應理論能否被成功的應用，關鍵之一取決於每個試題參數的準確性。理論上，要產生準確的試題參數，應在大樣本量（若使用二或三個參數對數型模式必須使用 1,000 位受測者；若使用一個參數，則可減少到 500 位受測者）下進行，然而大樣本量

在實作上卻因為獨特的樣本、試題的曝光率及有限的預算等因素考量，難以有效被執行。

一般認為，受測者依據其能力給予相對應的試題將可大大提升測驗時的效率(Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Stocking, 1990，引述自 Wu, 2010)，但其前提為題庫裡的試題，必須是建立在同一量尺上才行，否則試題間無法比較或延用。基於此，發表者提出電腦適性化預試（Computerized Adaptive Pre-Test , CAPT）方法，即在預試階段，如何給定每位受測者「最佳試題（Optimum Items）」，其主要原理說明如下：

(一)試題被主觀的區分為 5 個難度等級。

(二)每位受測者均包含 4 個相同試題。

(三)第一個試題是由難易程度中隨機挑選，其中難易程度取決於相同試題的原始分數。

(四)如果受測者通過更多的試題，則下一題就挑選難度較高的；反之，則挑選難度較低的試題。

另外，為證明所提出之 CAPT 方法是可行的，其舉出 2 類模擬研究數據，第 1 種是試題估算的準確性在常態分布（Normal Distribution）與均勻分布（Uniform Distribution）二者間之比較，第 2 種是試題與能力估算的準確性在電腦適性化預試（CAPT）與定錨不等組（Non-Equivalent groups with Anchor Test, NEAT）設計二者間之比較。從研究數據得出結論如下：

(一)試題估算的準確性，使用 CAPT 設計優於 NEAT 設計。

(二)在使用 CAPT 設計時，建議採用均勻分佈的能力。

(三)使用CAPT設計時，其難易度判斷的正確性是相當重要，也就是建議真實與主觀的判斷彼此是屬於中度相關性。

二、電腦化測驗發展趨勢

隨著資訊科技的發展、推廣及應用，過往最普及的紙筆測驗（Paper Based Tests, PBT）已逐漸朝向電腦化測驗（Computer Based Tests, CBT）發展之趨勢。而電腦化測驗的應用，則隨著使用時機與採取的立論不同，又可分為傳統的電腦化測驗（CBT）、電腦化適性測驗（CAT）及網路化測驗（WBT）（見表 5-1、電腦化測驗分類表）。

表5-1、電腦化測驗分類表

	傳統的電腦化測驗	電腦化適性測驗	網路化測驗
英文全名	Computer Based Tests	Computerized Adaptive Tests	Web Based Tests
英文簡稱	CBT	CAT	WBT
意義	將傳統紙筆測驗改成以電腦螢幕作為呈現介面，逐一或全部呈現試題的電腦輔助施測方式	針對不同能力程度的受測者及其作答速度，提供適合其能力作答的適當難度試題，以求估計受測者能力的最大精確性	就測驗環境而言，結合網際網路的優點，以提供超越時空、隨選隨測、高彈性施測環境的測驗方式
理論基礎	古典測驗理論	試題反應理論	古典測驗理論/試題反應理論
特色	1. 與傳統紙筆測驗內容相同 2. 施測及計分利用電腦輔助	1. 量身訂製的施測內容 2. 可顯現個別的能力差異 3. 降低考試挫折感	1. 網際網路的施測環境 2. 施測時間、地點彈性

		4.施測流程非線性 5.無法跳答 6.施測長度不同	
實例	International Computer Driving License：ICDL、Institute of Certified Management Accountants：ICMA	TOEFL－CBT、GRE、SAT	TOEFL-IBT

一般而言，傳統的電腦化測驗（CBT）是將傳統紙筆測驗（如選擇題、是非題）轉移到電腦上，讓受測者透過電腦螢幕閱讀試題，利用鍵盤或滑鼠移動游標點選答案。電腦化適性測驗（CAT）則是依據試題反應理論（Item Response Theory，IRT）為基礎所建置而成的電腦化測驗，該理論基本假設是根據受測者不同能力，給予不同的測驗題目，評量其實能力。網際網路（Internet）技術純熟與被廣泛使用，原本只能運作在個人電腦或區域網路上的電腦化測驗系統，逐漸被轉移到網際網路使用。以 Web 為介面，受測者經由瀏覽器（Browser）就可以輕易在世界各地隨時隨地進行測驗，比起早期只建構在個人電腦或區域網路的電腦測驗，在使用上更具彈性，而這類電腦測驗形式可稱為網路化測驗（Web Based Tests, WBT）或線上測驗（online testing）。

國際間有關電腦化測驗之演進，早在 1990 年代初期，美國 ETS（Educational Testing Services）即著手推動電腦化測驗（CBT），其中 GRE（Graduate Record Examination）測驗，在 1992 年即以電腦版本進行考試，1993 年更以 CAT 的形式，加入人工智慧自動調整試題，施行適性化評量，並於 1999 年全面停止紙筆型式的考試。另托福（TOEFL, Test of English as a Foreign Language）電腦版測驗於 1998

年 7 月起開始實施，而臺灣也於 2000 年 10 月開始採用電腦化適性測驗（TOEFL－CBT），2006 年 5 月起 ETS 再度將 TOEFL－CBT 改制為網路化測驗，實施新托福（TOEFL－IBT）後，即停用舊有之電腦托福。

總之，測驗的形式已從紙筆測驗逐漸發展至電腦化測驗，再至網路化測驗，而電腦化測驗所使用的媒介，也不再僅只是個人電腦，而可以是掌上型電腦（PDA）、電視機、手機等，受測者只要在任和地方就可連進來考試，而監試人員也只要從遠端來控制、管理、監試考試的進行即可。

三、電腦化測驗發展準則與標準

隨著國情差異與主客觀環境之不同，為了便於世界各地均能有所遵循電腦化測驗之發展，以促進跨越國界的一致性和基準之目的，於是有關電腦化測驗發展準則或標準便一一被制定，本次會議共介紹五項準則與標準，說明如下：

（一）ISO9126: 1991

ISO9126 為一項軟體品質的標準，提供客觀判斷一個系統是否達到品質要求與其所需的衡量方法。其所定義的品質模型，包含了可用性、可靠性、效率性、功能性、可移植性及可維護性等六項特性，而其中前三項則在 2002 年被用來作為審查 CBT 系統的評價依據。

（二）ATP Guidelines for Computer-Based Testing 2002

ATP（Association of Test Publishers）成立於 1992 年，為測驗與評量工具等服務進行評估、選擇、篩選、認證、教育或臨床用途的非營利性組織，其所制定出的該項指引主要針對高風險的測驗環境來進行規範，共分為規劃與設計、測驗的發展、測驗的管理、計分與計分報導、心理層面分析、利益相關者的連結等六大章節來說明如何有效發展電腦化測驗。

（三）BS7988: 2002

為英國標準協會（BSI）在 2002 年所公佈的一項標準，主要是教導如何運用資訊科技（IT）提供服務評量的指引與作法，目的是作為大學、學院、機構、法人及其他組織在提供各項考試服務時應注意的準則。

該標準說明組織在執行考試時，需要有效的方法來監視線上的受測者的行為，包含是否從螢幕採取強制性休息、防止螢幕偷窺、考試作弊、發送電子郵件及瀏覽網頁等，總共包含 17 個章節。

（四）ISO/IEC 23988: 2007

ISO/IEC 23988 為資訊科技評鑑作業（information technology- A code of practice for the use of information technology (IT) in the delivery of assessments）之國際標準，由於資訊科技已在教育等領域被廣泛用作測驗與評鑑的計分與記錄反應，若被適當地使用，將具有速度、效率、較佳的回饋等優勢，並可有效改進有效性與可靠性，但相對的，資訊科技也可能衍生出安全與公平等議題。

本標準範圍主要著重在三方面：適用的評鑑類型、適用的生命週期階段及具體的資訊科技問題，它適用於知識、理解及技能（即成就測驗）等評鑑，但卻不適用於職業、健康、能力與人格心理測驗等。

（五）International Guidelines on Computer-Based and Internet Delivered Testing:2005

該指導方針是 ITC 在 2005 年 7 月正式通過，主要建議在發展 CBT/Internet 時，首先應考量在施測過程中是否有監試人員參與，以及是否在特定場所進行測驗等情境（見表 5-2、測驗之管理模式）。在決定出所採用之管理模式後，便可從技術（Technology）、品質（Quality）、管控（Control）及安全（Security）等四個層面，找尋合理的解決策略，發展出適合的 CBT/Internet 測驗系統。有關技術、品質、管控及安全層面所應考量之要素，說明如下：

1.技術

主要是確保各項的軟硬體需求均被充分考慮，包含科技發展趨勢、CBT/Internet 優勢、科技對施測過程中所帶來的衝擊與影響、人為因素、特殊身分使用者之合理調整、應用的作業流程、提供受測者輔助訊息及練習操作機制等。

2.品質

主要是保證整體測驗過程的品質，包含適切的採用CBT/Internet、考量心理計量品質、由紙筆測驗改為電腦化測驗是否維持相同品質、測驗結果分析與分析結果正確性、測驗結果的解釋並提供適當回饋、各種受測群組的公平性等。

3.控制

主要是在測驗過程中，應提供適當程度的控制措施，包含詳列測驗條件的的控制層級、詳述施測時監督的控制層級、考前練習與試題曝光的管理、受測者的身分確認與作弊情形等。

4.安全

在測驗中確保作出適當的安全防護措施，包含注意測驗內容的安全性、受測者的資料在網路上傳遞時之安全、受測者作答結果的保密等。

表5-2、測驗之管理模式表

開放（OPEN）	任何人都可以進行並完成測驗	無須經由使用者身分驗證，且無監試人員	不安全模式
控制（Controlled）	測驗時必須經由登入程序	使用者必須輸入帳號/密碼，且無監試人員	中度安全模式
監督（Supervised）	測驗時必須經由登入程序，但地點不安全	監試人員監控著使用者的登錄帳號與密碼	安全模式
管理（Managed）	在特地場所進行測驗（如測驗中心）	監試人員查核使用者的身分	安全模式

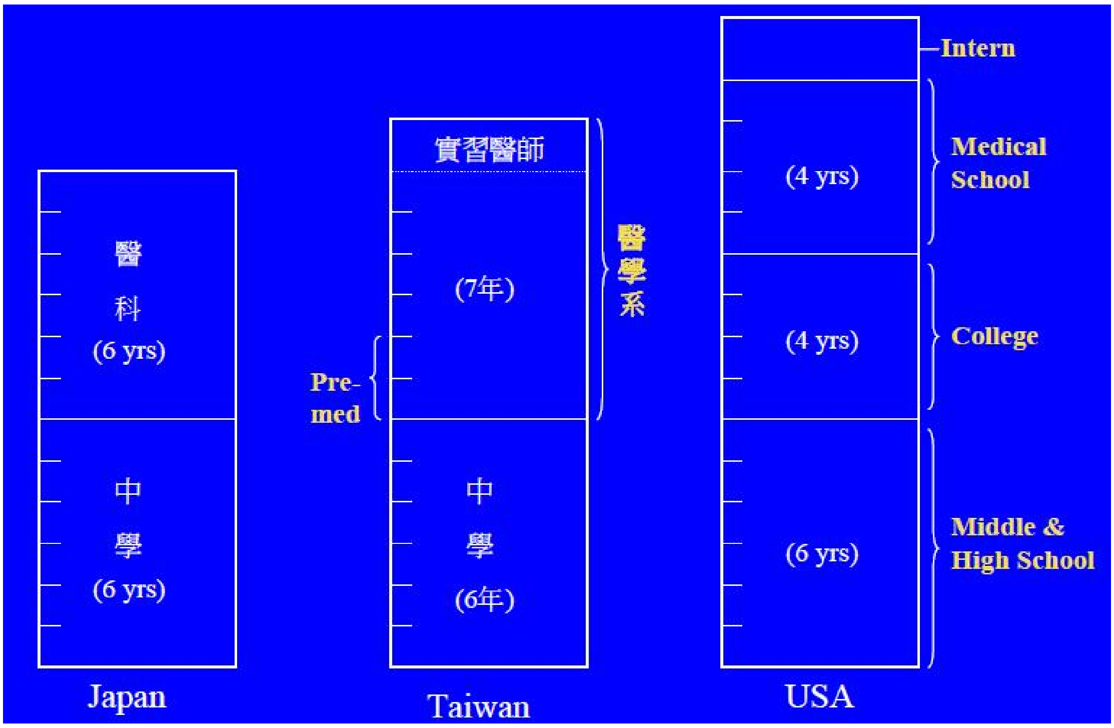
四、日本之電腦化測驗應用分享

世界各國對電腦化測驗之發展均有所著墨，而日本近幾年在電腦化測驗的應用上也有長足的發展，本報告僅就本次會議日本代表舉出該國電腦化測驗之應用案例與其困難點，說明如下：

(一)日本醫學臨床實習學校之共同成就測驗系統介紹

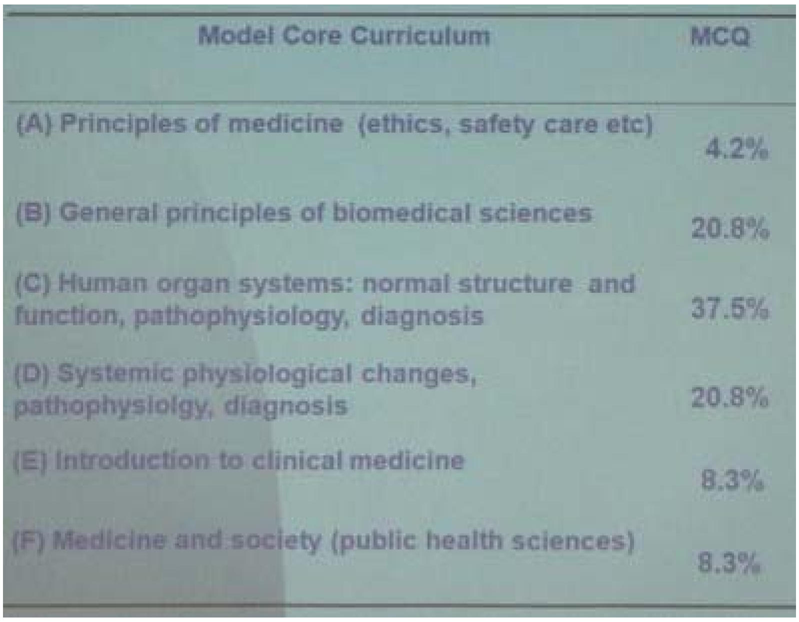
日本醫學系學制是高中畢業後便可直接進入醫學系就讀，課程修習年限原則是六年，前四年為基礎課程教育，後兩年則到醫院進行臨床實習（見表 5-3、日本與相關國家學制修習年限表）。為提高臨床訓練課程，該國在 2001 年 3 月提出核心課程（Model Core）概念，用以評量一項全國性的臨床技能與態度所需具備最低限度的基本知能。

表5-3、日本與相關國家學制修習年限表



另為有效衡量學生在臨床實習上的學習、表現及能力，在2001 年開始提出客觀結構化的臨床考試（Objective Structured Clinical Examination, OSCE）構想，並於 2002 年成立國家級的醫療共同成就性測驗機構（Common Achievement Test Organization, CATO），旨在改善該國醫療與牙科教育，以培育更多高素質的醫療人員。該機構除邀集全國醫學大學醫學院、牙醫學院學者外，並邀請醫科與牙科的學生，共同參與考題的命擬與製作，而所有考題題型均為複選題（Multiple Choice Question, MCQ）（核心課程與複選題分佈見表 5-4）。

表5-4、核心課程與複選題分布情形表



Model Core Curriculum	MCQ
(A) Principles of medicine (ethics, safety care etc)	4.2%
(B) General principles of biomedical sciences	20.8%
(C) Human organ systems: normal structure and function, pathophysiology, diagnosis	37.5%
(D) Systemic physiological changes, pathophysiolgy, diagnosis	20.8%
(E) Introduction to clinical medicine	8.3%
(F) Medicine and society (public health sciences)	8.3%

經過四年的努力，CATO 終於在 2006 年開始舉辦 OSCE 的電腦化測驗，有關該考試相關資訊說明如下：

1.考試試區：包含 80 所醫學與 29 所牙醫院校。

2.考試日期：依據各校課程時間辦理，原則為 12~3 月與 6~9 月。

3.考試人數：每所學校約 80~100 名。

4.考試題數：共 6 個科目（核心模型單元），總試題數為 360

題，其中 40 題為預測試題（pre-test items）。

5.考試題型：均為複選題。

6.考試內容：每位考生的試卷都由電腦經由亂數產生。

7.題庫總量：約 10,000 題。

8.試題命擬：每所學校每年提供 50~60 題，預估每年可成長約 4,000 題。

9.系統架構：採區域型網路之主/從式（Client/Server）架構（見圖 5-1、系統架構圖），試題由 CATO 將它存入光碟片/磁光碟片（CD/MO）後，再交付各試區匯入伺服器端電腦，以同時供應約 30~100 名考生上線考試，而考試結果則於考後統一由 CATO 寄發平均分數與個人分數成績單。

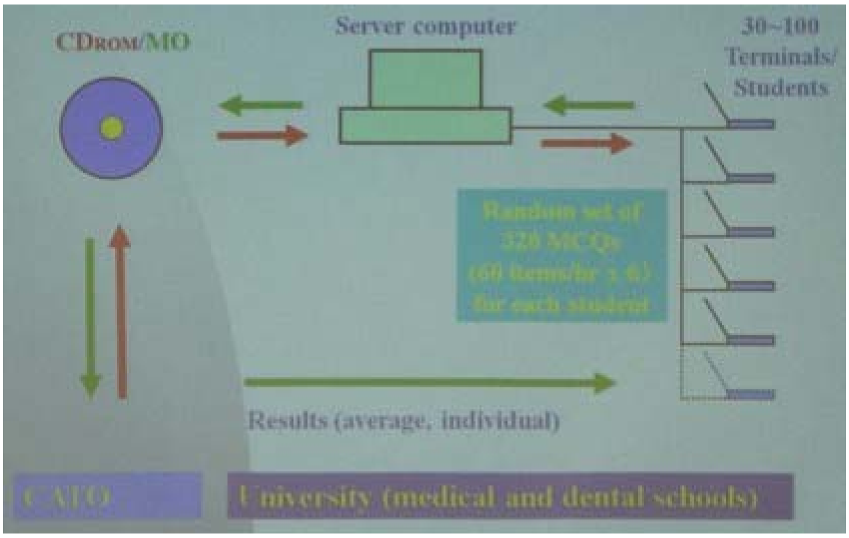


圖5-1、系統架構圖

(二)英語能力之電腦評量系統在日本發展與執行電腦化測驗的挑戰

在本場次會議（Nogami, 2010）中，日本代表開宗明義地說明了，該國考試制度身受中國古代的影響，因此其測驗文化具有下列幾項特性：

1.每年原則上舉辦一次考試，且考試時間是在同一天的同一

時間舉行，所有人的試卷內容均相同。

2.每次考試的考題都是全新的，也就是不重複考考古題或使用所謂的預測試題。

3.考試完後會公佈該次考試的試題內容。

4.不提供考生的分數報導結果。

5.命題者順便決定考試試題的呈現形式。

6.計分方式採原始分數（raw score），也就是從 0 分到滿分。

然而，為突破以上迷思，順應國際間之測驗發展潮流，該國於 2001 年 10 月成立了教育測量研究所（Japan Institute for Educational Measurement, JIEM），用以發展、通過、傳播良好的測量技術、測驗方法及測驗結果，以促進個人技能的發展。

JIEM 於 2002 年 5 月發展英語溝通的電腦評量系統（Computerized Assessment System for English Communication, CASEC），截至 2010 年 3 月，約有 7,600 萬人使用該項考試。該機構並陸續於 2005 年 11 月發展英語寫作自動評分系統（CASEC-Grammar, CASE-G）、2006 年 4 月發展（CASEC-G Tutoring System, CASE-GTS），未來並將陸續推出英文寫作導師（CASEC-Writing Teacher, CASEC-WT）、兒童英檢線上版（STEP Test for Children）等。

其中 CASEC 是基於試題反應理論（Item Response Theory, IRT）所發展的電腦化適性測驗（Computerized Adaptive Tests, CAT），期望在較短的時間內，可以準確地評估受測者的英文口語能力（圖 5-2、CAT 測驗原理），考試共分為詞彙、片語、聽力及聽寫四個科目（圖 5-3、CASEC 考試科目）。然而 CASEC 測驗在實務執行上，往往發生同一個試題出現在報考兩次考生身上的不平衡曝光頻率問題，為解決此類問題，他們採用了 Nogami(2010); Nogami, Kataoka & Mayekawa(2009, 2008)所提出的 3 參數轉為 2 參數理論，並且使用更有效的試題校準方法，如

固定較低漸近參數合理值、高效率的先驗分佈、小樣本條件試題校準（Tarre & Hong (2010); Baldwin (2006)）。

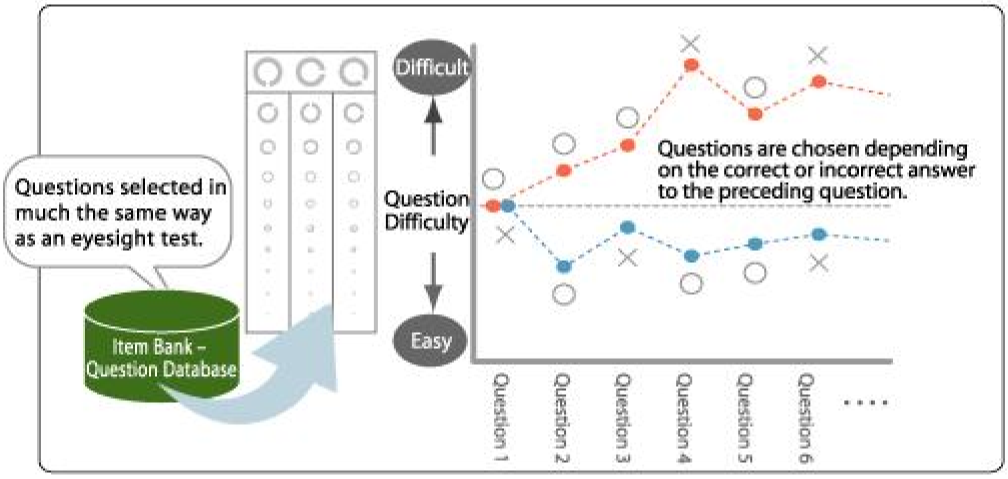
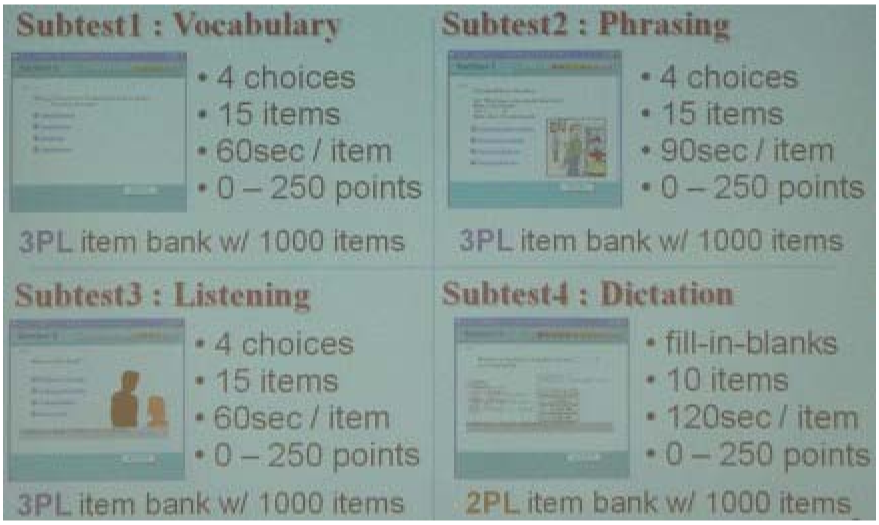

Questions selected in much the same way as an eyesight test.
Item Bank - Question Database
Difficult
Question Difficulty
Easy
Questions are chosen depending on the correct or incorrect answer to the preceding question.
Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
....

圖5-2、CAT測驗原理



Subtest1 : Vocabulary

- 4 choices
- 15 items
- 60sec / item
- 0 – 250 points

3PL item bank w/ 1000 items

Subtest2 : Phrasing

- 4 choices
- 15 items
- 90sec / item
- 0 – 250 points

3PL item bank w/ 1000 items

Subtest3 : Listening

- 4 choices
- 15 items
- 60sec / item
- 0 – 250 points

3PL item bank w/ 1000 items

Subtest4 : Dictation

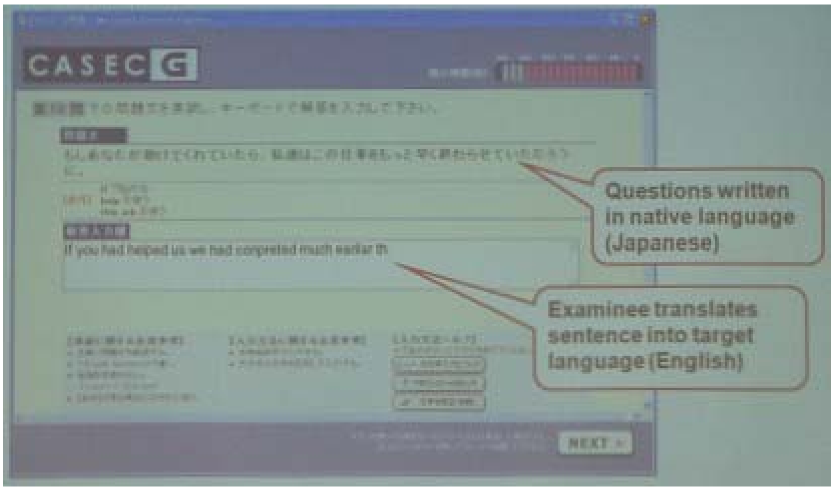
- fill-in-blanks
- 10 items
- 120sec / item
- 0 – 250 points

2PL item bank w/ 1000 items

圖5-3、CASEC考試科目

而 CASEC-G 與 CASEC-GTS 的發展目標則是幫助日本英文能力較低的學習者，透過自我學習的方式，提升英文寫作技巧。試想若從學習者開始練習英文作文寫作，交由老師進行批改，依據批改後之結果，學習者再重寫一次，如此反覆進行 2 至 3 次，雖是改善英文寫作技巧最常見的方法，但實務上卻難以執行。因

此 JIEM 提出了可以把重點放在如何將單一句子從母語（日文）翻譯到外國語言（如英文）的翻譯技巧上，有關該系統作答與回饋畫面範例，分見圖 5-4、CASEC-G 作答與回饋畫面範例與圖 5-5、CASEC-G 回饋畫面範例。

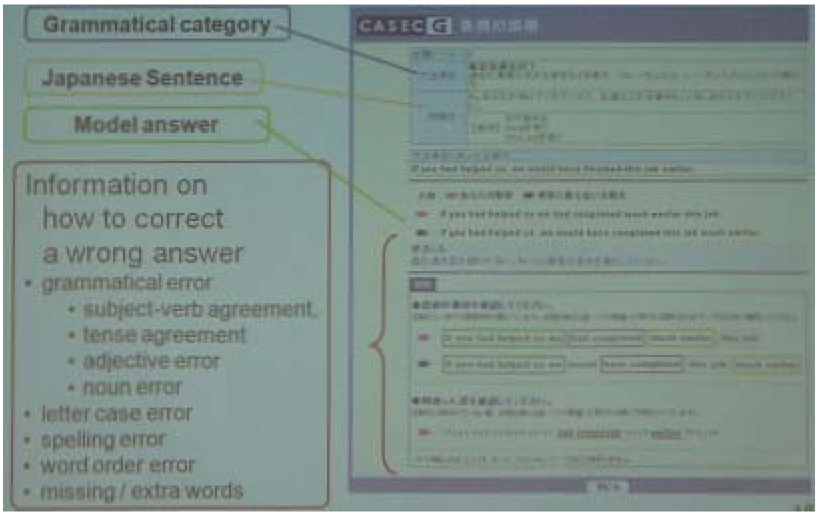


If you had helped us we had conpreted much earlier th

Questions written in native language (Japanese)

Examinee translates sentence into target language (English)

圖5-4、CASEC-G作答與回饋畫面範例



Grammatical category

Japanese Sentence

Model answer

Information on how to correct a wrong answer
- grammatical error
  - subject-verb agreement
  - tense agreement
  - adjective error
  - noun error
- letter case error
- spelling error
- word order error
- missing / extra words

圖5-5、CASEC-G回饋畫面範例

最後，JIEM 提出未來仍需面臨的挑戰如下：

1.網際網路：如隨時隨地可提供考試、雲端計算、智慧型手機等。

2.口語考試：為一種互動式的評量。

3.互動式學習系統：需考量處理程序與運作效能等問題。

（三）使用個人 IC 播放器之全國性聽力理解測驗

日本全國大學入學考試中心（National Center for University Entrance Examination, NCUEE）成立於 1977 年，為一獨立行政法人，負責管理日本全國高校招生考試中心與法學院入學考試。其所舉辦的大學入學考試（Nation Center Test, NCT）資訊如下：

1.考試時間：每年 1 月。

2.考試人數：每年約有 50 幾萬人報考（每年報考人數見表 5-5）。

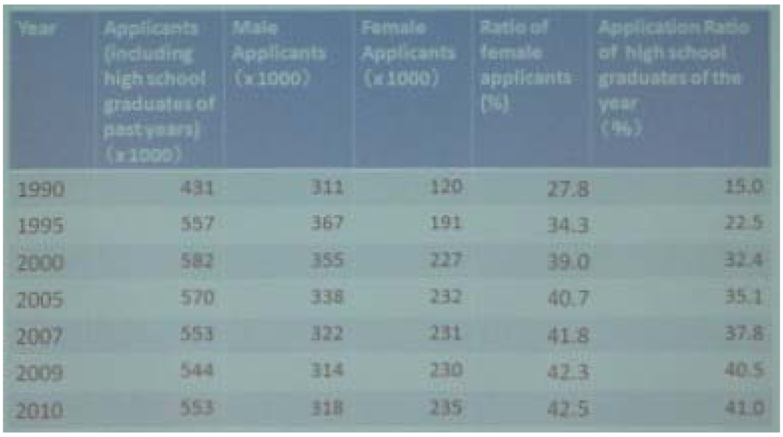
3.考試時間：為期 2 天，共 9 節。

4.考試題型：均為複選題。

5.未使用預試來進行試題難易度的等化。

6.考後立即公佈試題與解答。

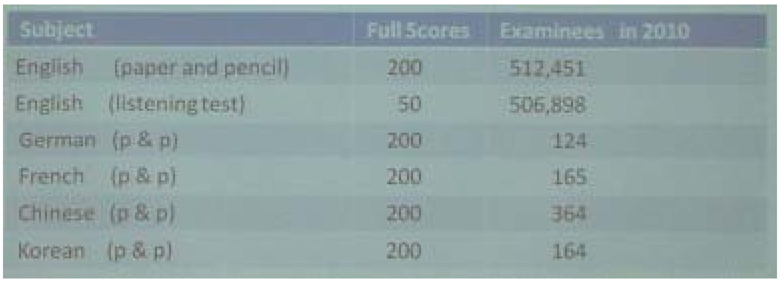
表5-5、NCT歷年報名人數統計表



Year	Applicants (including high school graduates of past years) （x1000）	Male Applicants （x1000）	Female Applicants （x1000）	Ratio of female applicants (%)	Application Ratio of high school graduates of the year （%）
1990	431	311	120	27.8	15.0
1995	557	367	191	34.3	22.5
2000	582	355	227	39.0	32.4
2005	570	338	232	40.7	35.1
2007	553	322	231	41.8	37.8
2009	544	314	230	42.3	40.5
2010	553	318	235	42.5	41.0

在 NCT 考試中，其中包含一門外國語科目，並提供英文、德文、法文、中文及韓文 5 種語文供選擇，絕大多數報考者均選擇英文，且只有英文才提供聽力理解測驗，其餘均採用紙筆測驗（NCT 外國語人數統計表見表 5-6）。

表5-6、NCT外國語報名人數統計表



Subject	Full Scores	Examinees in 2010
English (paper and pencil)	200	512,451
English (listening test)	50	506,898
German (p & p)	200	124
French (p & p)	200	165
Chinese (p & p)	200	364
Korean (p & p)	200	164

NCT 是在 2006 年才提供英文聽力理解測驗，所花費的時間共 1 個小時，前 30 分鐘為準備時間（包含講解、發答案卡、發設備、耳機設定、記憶卡設定、設備操作練習等），後 30 分鐘才是正式考試時間，所使用的測驗工具從 2010 年開始，統一由考試中心發放 IC 播放器（見圖 5-6、IC 播放器），考完後再回收重複使用。

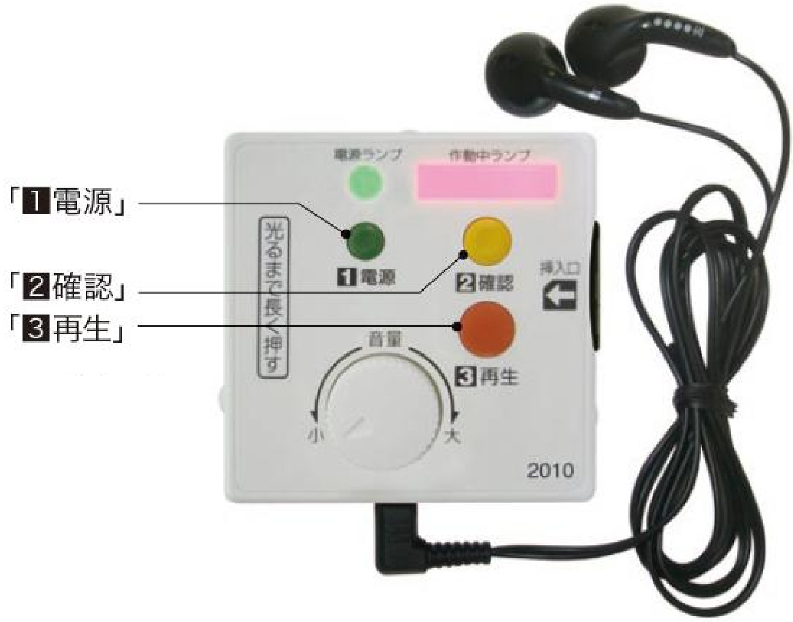


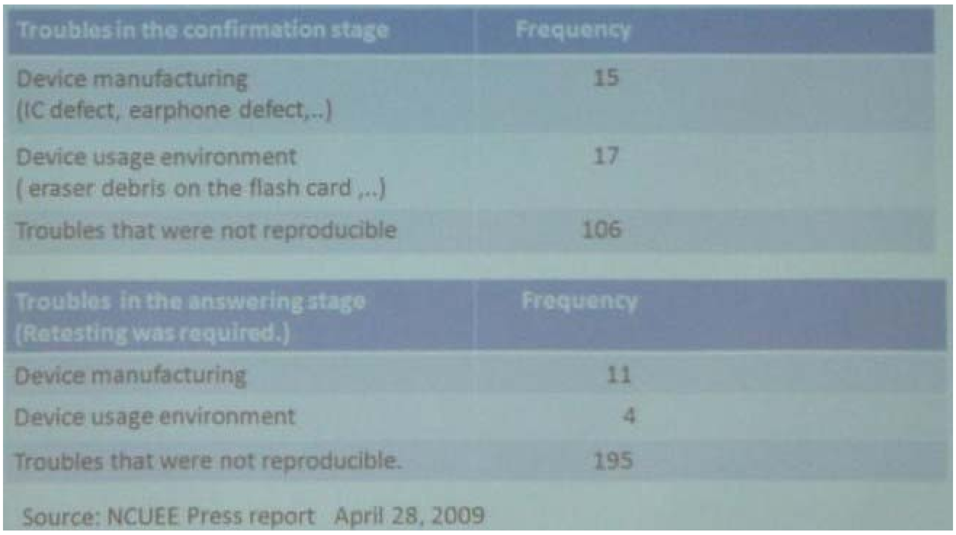
圖5-6、IC播放器

然而 IC 播放器存在著可能故障的風險（見表 5-7、設備故障原因與次數），因此 NCT 考試針對該設備發生故障時，明確規範其處理程序如下：

1.設備確認階段：立即更換新設備。

2.正式考試階段：安靜的請受測者填寫故障排除表格，並回收問題手冊、設備及答案卡，之後再另外安排時間從未完成之題目開始進行測驗。

表5-7、設備故障原因與次數表



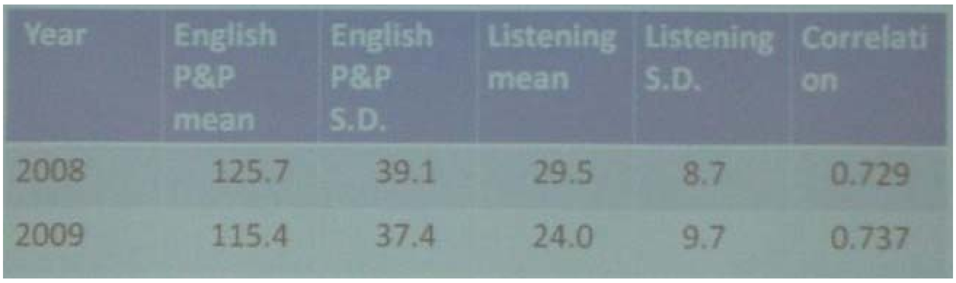
Troubles in the confirmation stage	Frequency
Device manufacturing (IC defect, earphone defect,..)	15
Device usage environment (eraser debris on the flash card ,..)	17
Troubles that were not reproducible	106

Troubles in the answering stage (Retesting was required.)	Frequency
Device manufacturing	11
Device usage environment	4
Troubles that were not reproducible.	195

Source: NCUEE Press report April 28, 2009

NCT 施行英文聽力理解測驗已 5 年，雖然不論是聲音的品質或與紙筆測驗的結果相比（見表 5-8、聽力與紙筆測驗結果統計表），均符合預期之目標，然而仍有設備成本費用、監考人員之訓練（設備操作與排除）及因設備故障需延期考試等問題尚待解決。

表5-8、聽力與紙筆測驗結果統計表



Year	English P&P mean	English P&P S.D.	Listening mean	Listening S.D.	Correlation
2008	125.7	39.1	29.5	8.7	0.729
2009	115.4	37.4	24.0	9.7	0.737

陸、結語與建議

本次參加 2010 年香港國際測驗委員會年會，收穫良多，以下僅就報告所提之各現研究發展趨勢，與我國國家考試相關的部分，提出相關的綜合結語及建議。

一、建立本部臨時命題、題庫建置、抽題與試題分析與回饋的品管準則

歐美先進國家對於考試整體的流程的進行品管準則的建立相當重視，包括測驗的管理、計分、編製試題、試題分析、測驗解釋、分數結果的報告、對測驗編製者與使用者進行訓練和督導、以電腦化資料管理系統處理測驗資料、制定測驗使用政策與測驗的出版均有標準化與科學化的處理程序。我國國家考試雖有訂定各種法規，但因行政人員測驗評量專業能力不足，命題常因受限於地域平衡、公私立校際平衡、一年內命題次數不能過多等等限制下，試題的品質穩定度並不容易掌控。本部又常常限於命題是一「高度屬人性」的工作，很少藉由嚴謹品管程序的規範影響委員的命題，雖命題委員的學科專業本部需予以尊重，但建立各項提升試題品質與確保信、效度的品管準則，本部責無旁貸。以下就與試題相關的品管準則建立方向整理如下：

（一）臨時命題的命題委員選擇，所命擬之應試科目試題，需確認是其主要研究領域與專長。

（二）無論是測驗題或申論題的臨時命題，均需落實配合本部公布之命題大綱各領域命題比例，利用「雙向細目表」命題，確保其「內容效度」與各領域試題之平衡。

（三）題庫建置對於各類試題的編碼、運用與回饋均應有詳細的紀錄，以了解命題委員的命題品質與使用效果。

（四）題庫建置常以命題委員分配命題題數的方式進行，但無論

是臨時抽題或電腦抽題，常無法完全配合該抽題科目的命題比例，建議未來抽題的原則，應考量各領域配分比例與難易度適當分配的原則。

（五）本部舉行的國家考試因非適性測驗，試題又需公布，優良試題之耗費相當可惜，每年命題為了不與三年內試題重複，許多命題反而流於冷辟，若能善用試題品質分析結果作為未來命題的依據，將能提升試題品質。

二、重視每次測驗結果報告作為下一次辦理考試的參據

目前國家考試之報名、試題反應與分數計算已進入電腦化的時代，但本部限於人力，很少針對測驗的各項統計數據作分析。先進國家的考試每一次測驗的結束，都會對測驗的各項統計數字與考試結果作分析，例如應考人背景、人數、考試次數、及格人數與及格率、錄取者之科系分布、及格率穩定情況……等等，但各項考試的測驗成果其實含有許多訊息可以供作考試單位參考，例如錄取不足額的類科、及格率偏低、考試人數驟增或驟降、試題難度偏低或偏高，對於檢討改進未來考試之辦理程序、提高國家考試的評量效能，都有極大的幫助。

三、提供應考人成績單完整的計分訊息

目前應考人參加國家考試後，測驗題經電腦讀卡，申論題經評閱後，各科分數經查核經電腦完成加總後，印出紙本成績單寄送，惟應考人成績單所呈現之計分資訊過於簡單，僅有各科原始分數及總分以及錄取與否之資訊，無法使應考人深入了解各科真實表現。近年來參加國家考試應考人呈現遽增之現象，大多數之應考人非參加一次國家考試即能榮登金榜，提供全部應試科目及總成績計分資訊，除能讓應考人了解各科之優、缺點，提供未來之準備方向與重點，更能彰顯本部為應考人設想之用心。

國外大型測驗機構早已進行應考人各科分數與總成績詳實的報告作業，以提供應考人回饋訊息。建議除供提供申論題的題分數據外，可藉由應考人所熟悉的基本統計概念「百分等級」(PR 值）以及「長條圖」的輔助，在成績單上表示應考人各科之相對位置，以利應考人了解各科之表現狀況。

四、善用多元評量方法提升國家考試評量效度

目前的國家考試的筆試大多屬「分科成就測驗」，僅有民航人員考試民飛航管制人員委由航醫中心進行「心理性向測驗」，評量與檢查的項目包括動作判斷能力、認知及推理能力、空間能力，因民航管制人員的工作與飛航安全息息相關，僅藉由學科專業知識的筆試難以甄拔適任人員。同樣的，在我國的公務人員考選中，與民航管制人員類似的考試至少有外交人員、警察人員、國安人員、調查人員….等，國外考選此類人員已普遍使用心理測驗，例如人格測驗與性向測驗，人格測驗對於篩選不適任人員的效用甚大；性向測驗則對於預測未來表現傑出的效用也極為顯著，本次國際測驗委員會所介紹的人格測驗與情緒智力測驗都有篩選適任人員的效果，但因編製此類測驗之專業性甚高，唯有先行確認各類國家考試特殊人員所需的工作職能，以及在實務工作中觀察不適任人員的反應與特質，編製具有區辨效度的心理測驗，測驗的構面例如情緒穩定度、抗壓型、外向性、心理調適力、積極性….等等。雖編製成本甚高，耗時甚長，但對於前述特殊公務人員的有效甄選，節省未來的訓練與社會成本，絕對是有價值的。

五、提升國家考試試題品質配套措施

國家考試對於絕大多數的應考人而言，是一項極具關鍵性之考試，因為它足以影響許多人的一生。因此本部在辦理各項國家考試時，除了應秉持著公平、公正、公開之基本原則外，就測驗的

角度而言，考試的內容更應具備良好的信度、效度及鑑別度。要有效提升國家考試的信度、效度及鑑別度，吾人認為可從幾方面著手：

(一)部分考試類科推行預試制度

國家考試基本上屬於成就測驗的一種，所謂成就測驗指的是，測量受測者經過學習或訓練之後所獲得的知識或能力，也就是可以透過成就測驗客觀的瞭解一個人在某學科上獲取多少知識，或者在綜合科目學習上的成就水準為何，以及與年級或年齡同儕差異的情形為何。要達成標準化的成就測驗，其中一項步驟即是預試與分析試題特徵，預試主要目的是進行試題的擇優汰劣與建立測驗的信、效度，因此預試結果的準確性也就相形重要，故在執行預試時應注意：

1.預試對象宜取自將來正式測驗擬應用的群體，並注意代表性。

2.預試實施應與正式測驗時情況相似。

3.預試時，應使受測者有足夠作答時間，以蒐集充分的反應資料，俾使統計結果更為可靠。

4.過程中應記錄受測者反應(不同時限內一般受測者所完成之題數、題意不清之處，及其他有關題目)。

然而預試必須花費相當多的人力與成本，在國家考試上千個應試科目當中，若要全部採行預試制度，恐怕難以有效實施，因此建議可先從歷年來試題疑義較高之類科（醫事類科）、爭議性較高之考試（司法官考試）或未來要推行電腦化適性測驗之考試優先進行，期望在本部有限資源下，逐步推行標準化測驗，以發揮考試公平、公正之最大價值。

（二）增列複選題型

本部現行國家考試試題題型共分申論式與測驗式 2 種題型，其中測驗試試題又為四選一之單選題，以測驗的觀點，實難有效

測出受測者之真實能力，故可考慮發展複選題型，並採倒扣計分方式，除可避免受測者猜題行為外，並可提高測驗的鑑別度。

六、改進報名程序同時需考量防弊措施

本部刻正積極推行網路單軌報名作業，並朝無紙化方式邁進，這是相當好的改革方向。但仍需在簡化報名流程，使用科技工具的同時，注意各項防弊措施。因為即使如歐美各國採行電腦化報名與電腦化適性測驗或網路測驗的國家，仍然相當重視個人正確身分的查核與個人生理與外在特徵的辨認與存檔。以往國家考試因為有應考人的相片與畢業證明文件作為確認真實身分的工具，雖然資料寄送繁瑣，但防弊功能甚大。尤其應考人的身分證明文件已經是民國 95 年左右核發，新式身分證的相片與目前面貌可能已差異甚大，對於監場人員查核正確身分有所困難。當然我們也不必像大陸在應考前一天需到考場拍照存檔，徒增應考人負擔。但在以往的國、內外各種考試中，假造畢業證明文件的案例層出不窮，若無嚴密的防弊措施，將給予投機之應考人鑽漏洞的機會。

為使報名程序提高效率，又能兼顧防弊措施，建議採取網路報名時仍需請應考人繳交六個月內的二吋照片電子檔。至於畢業證書繳交，國內已使用電腦化建立畢業生資料的學校，當然可藉由網路連線查核；其餘國外畢業生或離開學校太久的畢業生，仍應繳交紙本或掃描成電子檔傳送至考選部。在簡化繳交證明文件程序時，參加第一次國家考試者的畢業證書與證明文件就應先行存檔，若是下次再參加國家考試時，就不用繳交畢業證書，直接從資料庫調檔查核即可，是相當便民的行政措施。

七、提昇電腦化測驗評量效能

(一)參酌國際標準規劃

由於國情與環境等差異，各國乃至於各種測驗是否要運用電

腦來進行評量，都有其需特別考量之處，如測驗的目的、高風險（high stakes）或低風險（low stakes）、測驗過程是否需監督（supervision）、線上或離線考試（online/offline）、使用光碟（CD）或下載可執行文件（download executable）、全採電腦測驗（fully computerised testing）或部分電腦測驗（part-computerised testing）等。

本部電腦化測驗系統於民國93年完成client-server第一版並實施6年後，於99年再完成Web版之更版作業並開始實施，期間雖已累積相當多之系統開發與考試實施時之經驗，但未來在進行系統規劃與開發過程中，除了可持續安排考察或參加國際型會議，以觀摩國內外各項成功案例外，若能適時參照相關國際標準，完整考慮各種測驗與系統規劃面項，如測驗的發展、適用的評鑑類型、測驗的管理、計分與計分報導、心理層面分析、利益相關者的連結、系統之可用性/可靠性/效率性/安全性等，必能有效發揮系統最大服務價值。

(二)精進測驗內容與作答型式

本部電腦化測驗已施行6年多，並具一定之發展規模，然其令人詬病的原因主要有2項，一是系統建置花費過高，而使用率卻偏低，因此如何擴大範圍辦理，是本部一直在思考改進的，例如本部預定在100年增列物理治療師等類科，並配合增建1,700席電腦試場；而另一項令人詬病的是，電腦化測驗僅將傳統紙筆考試轉移至電腦作答，未發揮出系統真正的價值。

有關這部分，吾人認為本部在持續擴大辦理施測範圍之際，可同時致力於測驗實質內容之提升，如多媒體試題、聽力測驗、互動式測驗、情境式測驗等，以發揮紙筆測驗所不能及之處，並加速推動申論題、繪圖題等電腦作答介面，整合後續人工電腦閱卷作業，除提供應考人較熟悉且便利之作答介面外，並可有效提升評分信度，開創更多資訊科技所帶來之加值服務。

(三)發展電腦化適性測驗

世界各國已有諸多施行電腦化適性測驗之成功案例，如TOEFL－CBT、GRE、SAT，因此只要決定出適合之選題規則、受測者能力估算法、選題策略、分數等化、試題曝光率與重複使用率、考試終止規則等內容，相信發展電腦化適性測驗應無技術上之問題。

因此，吾人認為可先從專門職業及技術人員高等暨普通考試醫事人員之藥師類科考試優先辦理電腦化適性測驗，主要係考量每次報考藥師人數約2千多人，符合本部電腦化測驗試場座位規模，且該項考試每年共舉辦2次，施行電腦化適性測驗所遭受之衝擊與困難相對較低，同時可提高評量信、效度。

然而，因本部現行國家考試制度與我國國情，在發展電腦化適性測驗之前，應先設法解決考後不公布試題等政策性問題，並且擬定電腦化適性測驗之試題研發流程、基本理念、採用之效益、考試之公正性等說帖，加強宣導，以免除社會大眾之疑慮。除此之外，電腦化適性測驗之試題必須經過試題預試之標準程序，以擇優汰劣試題，並客觀建立出試題之難易度，考量預試作業需花費大量之人力、物力、時間、費用等成本，故建議100年完成藥師類科之題庫試題建置作業（含預試），於101年正式施行該類科考試電腦化適性測驗，促使國家考試之電腦化測驗再邁向新的里程碑。

參考文獻

江文慈（2005）**何倫碼-Holland 生涯類型**，世新大學師資培育中心課程資料簡報檔。

金樹人（2009）**生涯諮商與輔導**，台北市：東華書局。郭生玉（2004）**教育測驗與評量**。台北市：精華書局。

歐慧敏（2006）情緒智力理論及其評量。**教育研究月刊**，147 期，146-155。

Avi Allalouf（2010）**Establishing the Guiselines on Quality Cntrol in Scoring,Analysis and Responding of Test** Scores.2010 Itc Commision Workshop,18,July 2010, Hong Knog.

Cost.P.T.,Jr.,& McCrae,R.（1989）.**NEO PI-R test manual.Port Huron,MI:**Sigma Assessment System.

De Fruypt,Filip（2010）.**The Big Five in Selection and development assessment：Blessing and yoke.** 2010 Itc Commision essay.

International Test Commsion（2000）.**International Guidelines for Test Use.**www.intestcom.org/itc.projects.htms

Schmit,Neal（2010）**Impact of Measurement Invarances on construct correlation,mean differences,and relationships with external correlates:Big Five and RIASEC Measures.**2010 Itc Commision essay.

Serlie,Alec、Hiemastra,Annenarie、Van Leeuwen,Rob,& Bazen,Madelijn（2010）.**Are there Big Culture differences in Personality？**2010 Itc Commision essay.

Wong,Chi-Sum、Peng,Kelly & Huang,Emily（2010）**Alternative Methods assessing the emotional intelligence of Chinese** respondant. 2010 Itc Commision essay.

Foster, David (2010), International high-stakes online testing: best practices for test security and data privacy. 2010 Itc Commision essay.

Fremer, John (2010), International trends in test security - Certification testing. 2010 Itc Commision essay.

Burke, Eugene, International trends in test security - Employment testing. 2010 Itc Commision essay.

B. F. Green (1991), Computer-Based Adaptive Testing in 1991, Psychology & Marketing, 8 (4).

Wu, Chia-Ju (2010), Specifying optimum items to the examinees for item parameter calibration in pretest: The computerized adaptive pretest. 2010 Itc Commision essay.

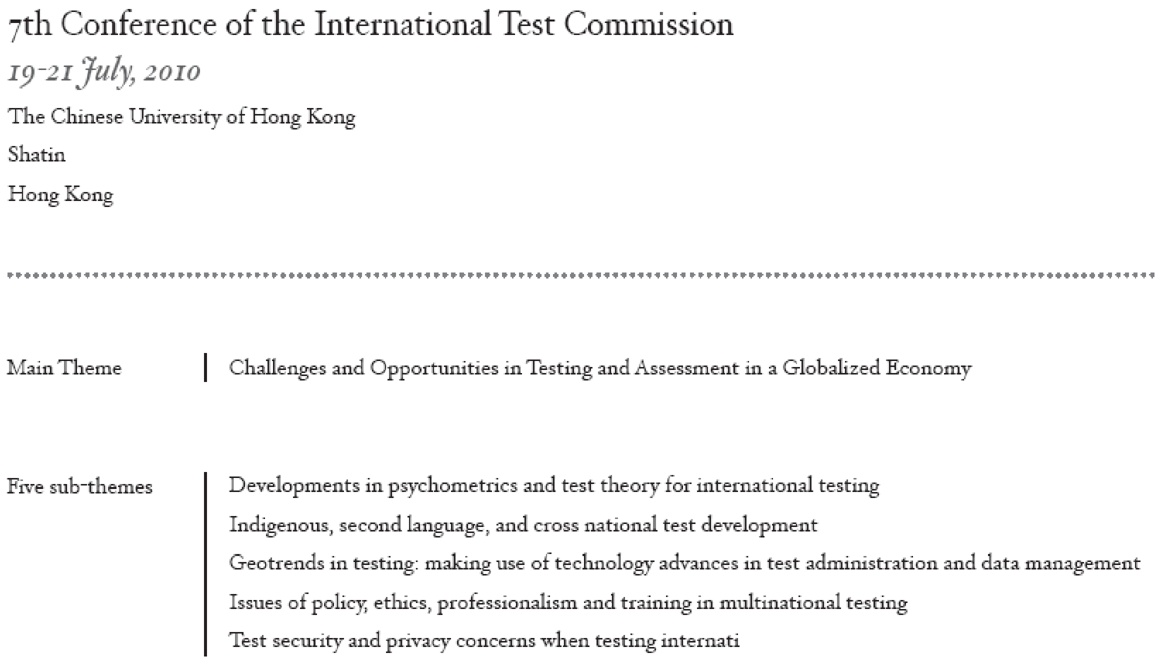
Mayekawa, Shinichi (2010), Introduction to the Common Achievement Test System for entering clinical clerkship in Japanese medical schools. 2010 Itc Commision essay.

Nogami, Yasuko (2010), Challenges in developing and operating CBTs in Japan. 2010 Itc Commision essay.

Otsu, Tatsuo (2010), A nationwide listening comprehension test using personal IC players. 2010 Itc Commision essay.

附錄一、第七屆國際測驗年會會議議程

Scientific Programme



7th Conference of the International Test Commission

19-21 July, 2010

The Chinese University of Hong Kong
Shatin
Hong Kong

Main Theme	Challenges and Opportunities in Testing and Assessment in a Globalized Economy
Five sub-themes	Developments in psychometrics and test theory for international testing Indigenous, second language, and cross national test development Geotrends in testing: making use of technology advances in test administration and data management Issues of policy, ethics, professionalism and training in multinational testing Test security and privacy concerns when testing internati

Programme Schedule

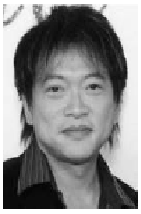
Time	LT1	LT2	LT3	LT4	LT5	LT6	Room 201	Room 211	Room 105
7:30-8:30	Registration (Level 2)								
8:30-9:15	Opening Ceremony (LT1)								
9:15-10:45	Plenary Session (LT1) State-of-the-art Speech by Robert Roe Speech by Dave Wilson from GMAC								
10:45-11:00	Tea Break								
11:00	*Keynote Address 1* ***Validation support for selection procedures*** ***Schmitt, Neal*** Weiss, Larry (session chair)	*Invited symposium 1* ***Allalouf, Avi*** Bennett, Randy Hambleton, Ronald Zhang, Houcan Grégoire, Jacques Gafni, Naomi		*C01* ***Leong, Frederick*** Arce-Ferrer, Alvaro Li, Xuhang Schulruf, Boaz Wu, Joseph Zumbo, Bruno D.	*Special Session 1* ***Debating the cost of psychological tests and the factors that determine the cost*** Foxcroft, Cheryl Bartram, Dave Foster, David Oakland, Tom	*Diamond Sponsor Session 1* ***Anastasio, Ernie*** Guo, Fanmin Talento-Miller, Eileen Defibaugh, Courtney Taliaferro, Hillary Rudner, Lawrence		*S01* ***Brown, Gavin T L*** Michaelides, Michalis Gao, Lingbiao Hui, Sammy K. F. Kennedy, Kerry J.	
11:30									
12:00									
12:30-14:00	Lunch								
14:00	*Keynote Address 2* ***Ethical and other professional issues: What to do when working in the absence of local standards*** ***Oakland, Thomas*** Purwono, Urip (session chair)	*C03* ***Talento-Miller, Eileen*** Clinton, Janet M. Zheng, Ying Johnston, Michael Orchard, Sue Zumbo, Bruno D. Michael, Joan J.		*C02* ***Leung, Kwok*** Chen, Lijun Eatchel, Nikki Vrignaud, Pierre Wang, Y. Lawrence Xu, Jian Ping	*S02* ***Weiner, John*** Burke, Eugene Fetzer, Michael	*Special Session 2* ***International Test Commission Guidelines for adapting educational and psychological tests (2nd edition)*** Bartram, David Gregoire, Jacques Hambleton, Ronald Van de Vijver, Fons	*S04* ***Zhou, Dan*** Li, Bo Li, Chongliang Li, Qi Xu, Jiehong		***P01*** Chao, Yu-Ning Gomiero, Tiziano Kreuzpointner, Ludwig Kuo, I-Ting Li, Ying Lu, Cheng-Chin Mitina, Olga Morley-Kirk, James Ni, Chen-Hao Preuss, Achim Roberts, Patricia Hill, Jill Gnambs, Timo Tono, Suwartono Hung, Pi-Hsia Sun, Jianan Wang, Bo
14:30									
15:00									
15:30-15:45	Tea Break								
15:45		*C04* ***Schwarzer, Ralf*** Cheung, Shu Fai Chiou, Hawjeng Mitina, Olga Yuan, Ke-Hai Zhang, Zhiyong	*C05* ***Wang, Wen Chung*** Albers, Frank Arce-Ferrer, Alvaro Chernyshenko, Oleksandr Huelmann, Gerrit Leung, Chi Keung Eddie Zierke, Oliver	*S06* ***Leung, Freedom Yiu Kin*** You, Jianing Lai, Ching Man Fu, Kei	*Invited Symposium 2* ***Nielsen, Sverre Leonard*** Lindsay, Geoff Oakland, Tom Bartram, Dave Hagås, Per Olav	*Invited Symposium 3* ***Weekers, Anke*** Schmitt, Neal De Fruyt, Filip Serlie, Alec Van De Vijver, Fons	*C06* ***Elosua, Paula*** Choi, Hye-Jeong Defibaugh, Courtney Solano-Flores, Guillermo Talento-Miller, Eileen Zheng, Ying	*S05* ***Harris, William*** Schachart, Nadine Tong, Alex Burke, Eugene	***P02*** Chen, Su-Yu Ding, Shu-Liang Fang, Ping Ghadimi Moghaddam, Malek Mohammad Kuo, Bor-Chen Kwok, Oi-Man Leeson, Heidi Leontiev, Dmitry Zhang, Minqiang Liu, Yan Lozano, Luis M. Lu, Saucheng Wang, Aijun Wang, Wen-Yi Xie, Qin Yang, Zhiming Yoon, Myeongsun Zheng, Rui Xin, Tao You, Xiaofeng
16:15									
16:45									
17:15-17:30	Break								
17:30-19:30	Welcome Reception (Level 3)								

Time	LT1	LT2	LT3	LT4	LT5	LT6	Room 201	Room 211	Room 105
8:00-8:30	Registration (Level 2)								
8:30			***S03*** ***Cheung, Kwok Wah*** Cheung, Kwong Yuen, ThomasLam, Ling Chi, TennyTang, Mei Shin Chan, Ka Ki, Catherine	***C08*** ***Roth, Hans*** Foxcroft, Cheryl Van De Vijver, Fons Yu, Guoxing Zhang, Sheng Hill, Jill	***S07*** ***Fan, Weiqiao*** ***Zhou, Mingjie*** Wan, Sarah Lai Yin Cao, Hui	***S08*** ***Chen, Po-Hsi*** Wu, Chia-Ju Chao, Hsiu-Yi Hsu, Chia-Ling Chen, Jyun-Hong	***C09*** ***Nielsen, Sverre*** Bartram, Dave Phelps, Richard Preuss, Achim Serlie, Alec W. Sharf, James	***C07*** ***Weekers, Anke*** Choi, Youn-Jeng Griffin, Patrick Kingston, Neal	***P03*** Coscarelli, Alessandra Barbot, Baptiste Chang, Kenneth Charalampous, Kyriakos De Bastiani, Elisa Ding, Ching-Huei Elosua, Paula Fida, Kashif M. Iversen, Ole Lam, Ben C. P. Laurentiu P., Maricuoiu Lee, Leanda Lozano, Luis M. Megherbi, Hakima Okonkwo, Judith So, Timothy Vrignaud, Pierre Wan, Shih-Ting Wu, Shiu-Lien Brown, Gavin Thomas Lumsden
9:00	***Keynote Address 3*** ***International high-stakes online testing: Best practices for test security and data privacy*** ***Foster, David*** Byrne, Barbara (session chair)	***Special Session 3*** ***The international Journal of Testing: Ten Years and Going Strong*** Hattie, John							
9:30									
10:00-10:30		Tea Break							
10:30		***S09*** ***Burke, Eugene*** ***Tong, Alex*** Liu, Ying	***S18*** ***Bernstine, Daniel O.*** Van Tol, Joan Vaseleck, James Pashley, Peter Rudner, Lawrence	***S12*** ***Gilmore, Alison*** Smith, Jeffrey K. Darr, Charles Smith, Lisa Hattie, John	***S10*** ***von Davier, Alina A.*** Gorham, Jerry Eggen, Theo Yoo, Hanwook Fetzer, Michael S. Gafni, Naomi		***C10*** ***Natarajan, Venkatesa*** Barrada, Juan Ramon Han, Kyung (Chris) T. Lin, Jyun-Ji Liu, Hongyun Walker, Cindy M.	***S11*** ***Ding, Yi*** Kuo, Yi-Lung	***P04*** Ertubey, Candan García-Rueda, Rebeca Gillet, Isabelle Kuo, Bor-Chen Hsu, Chun-Yu Lin, Yi-Hung Lozano, Luis M. Li, Tao Ding, Shu-Liang Molchanov, Alexander Okonkwo, Judith Primi, Caterina Rogers, H. Jane Shih, Pei-Chun Su, I-Hsiang Wang, Aijun Zhang, Minqiang Zhao, Yue
11:00	***Keynote Address 4*** ***From indigenous to cross-cultural personality assessment: The usefulness of the combined emic-etic approach*** ***Cheung, Fanny M.*** Leong, Frederick (session chair)					***Special Session 4*** ***Examining Formative Assessment*** Bennett, Randy Brown, Gavin Koh, Kim			
11:30									
12:00-13:30	Lunch								
13:30	***S19*** ***Fremer, John*** Burke, Eugene Geranpayeh, Ardeshir Tong, Alex Sun, James Jian-Min	***S13*** ***Wang, Wen Chung*** Lee, Kung-Hsien Huang, Sheng-Yun Chen, Po-Hsi	***S14*** ***Chen, Shu-Ying*** Chen, Jyun-Hong Lin, Yi-Hung Liu, Tzu-Chen Lee, Hsiang-Ling	***S15*** ***Gan, Yiqun*** Ng, Alexander Yao, Jingdan Fan, Weiqiao Leung, Kwok	***Invited Symposium 4*** ***Purwono, Urip*** Halim, Magdalena Mansyur, Mansyur Salim, Djohan Mardhapi, Djemari Suhapti, Retno	***Diamond Sponsor Session 2*** ***Schuchart, Nadine*** Gillet, Isabelle Li, Tao Roth, Hans J. Schuchart, Nadine	***C12*** ***Ercikan, Kadriye*** Egeland, Jens Hill, Jill Malykh, Sergey Mordeno, Imelu Yan, Gonggu	***C11*** ***Johnston, Michael*** Gessaroli, Marc Lee, John Chi-Kin Mamauag, Maria Felicitas (Marife) M. O'Neill, Thomas Woo, Ada	***P05*** Curkovic, Natalija Dela Rosa, Elmer Ganotice, Fraide Li, Jian Li, Zhongquan Turilova-Miščenko, Tatjana Zulu, Happy Amurao, Analiza Liezl Arce-Ferrer, Alvaro Ding, Shu-Liang Gorham, Jerry Han, Yuna Lee, Pei-Yu Lin, Chien-Yu Molchanov, Alexander Park, Yoon Soo Cheng, Chien-Ming Wang, Li Jun Ye, Shengquan Yu, Jiayuan Ong, Saw Lan
14:00									
14:30									
15:00-15:15	Tea Break								
15:15	***Invited Symposium 5*** ***Hambleton, Ronald K.*** Huff, Kristen Mills, Craig Zumbo, Bruno D.	***S16*** ***Mok, Magdalena Mo Ching*** ***Wang, Wen Chung*** Wang, Li-Jun Tam, Hak-Ping Cheng, Rebecca Wing-Yi Lee, Tony	***C13*** ***Han, Chris*** Barbot, Baptiste Care, Esther Chiou, Hawjeng Lee, Young-Sun Magno, Carlo Park, Yoon Soo	***Invited Symposium 7*** ***Shigemasu, Kazuo*** Muraki, Eiji Mayekawa, Shinich Nogami, Yasuko Otsu, Tatsuo	***Invited Symposium 6*** ***Fontaine, Johnny*** Wong, Chi-Sum Roberts, Richard Yik, Michelle Grégoire, Jacques	***S17*** ***Bartram, Dave*** Evers, Arne Born, Marise Sun, James Jian-Min	***C16*** ***Gilmore, Alison*** Alexeev, Natalia Ong, Saw Lan Phelps, Richard Shih, Shu-Chuan Syaifuddin, M.	***C14*** ***Leung, Freedom*** Bittner, Jenny V. Dasari, Venkata Venu Gopal Gori, Alessio Leung, Cynthia Tao, Vivienne Y. K. Shahid, Mamoona	***P06*** Besharat, Mohammad Ali Chang, Te-Sheng García Meraz, Melissa Hanfstingl, Barbara Hatami, Mohammad Hou, Ya-Ling Huang, Duan Hung, Su-Pin King, Ronnel Chang, Kuei-Lin Li, Xueyan Michaelides, Michalis Pariñas, Neil Vrignaud, Pierre Xu, Jian Ping Yang, Min Nacira, Zellal Zhang, Min-qiang
15:45									
16:15									
16:45			***C15*** ***Chang, Lei*** Ambreen, Saima Bertling, Jonas Pablo Jeng, Hi-Lian Magno, Carlo Marberger, Tove Kanestrom	***S20*** ***Cowieson, Neil*** Fraccaro, Michael U, Kin Chong Fung, Helen To, Clara			***C17*** ***Mok, Magdalena Mo Ching*** Yang, Zhiming Han, Kyung (Chris) T. Kubinger, Klaus D Fung, Tze-Ho Huang, Hung-Yu Chernyshenko, Oleksandr		
17:15						***AGM of ITC and Town Hall Meeting***			
17:45									
18:15-18:30									
18:30-21:00	Banquet (The Star Seafood Floating Restaurant)								

Time	LT1	LT2	LT3	LT4	LT5	LT6	Room 201	Room 211	Room 105
8:00-8:30	Registration (Level 2)								
8:30		***S21*** ***To, Clara*** Bartram, Dave Elliott, Ray Sue-Chan, Christina Leung, Kwok	***S22*** ***Cheung, Shu Fai*** ***Born, Marise*** Leong, Frederick Iliescu, Dragos Dang, Minh	***C19*** ***Oakland, Tom*** Cheng, Christopher H K Freund, Alexander Mcinerney, Dennis Sava, Florin A. Hill, Jill Van Luijk, Frank	***Invited Symposium 8*** ***Weiss, Lawrence G.*** Chen, Hsin-Yi Li, Yuqiu Zhang, Houcan Zou, Yizhuang Grégoire, Jacques	***Invited Symposium 9*** ***Elosua, Paula*** ***Hambleton, Ronald K.*** De Boeck, Paul A.L. Elosua, Paula Zumbo, Bruno D.	***S23*** ***Burke, Eugene*** Harris, William G. Rudner, Lawrence Sun, James Jian-Min Geisinger, Kurt. F. Foster, David	***C18*** ***Weekers, Anke*** Dodeen, Hamzeh Zheng, Ying Kuo, Bor-Chen Swaminathan, Hariharan Wehrmaker, Maike	***P07*** Besharat, Mohammad Ali Curkovic, Natalija De Bastiani, Elisa Fan, Xiao Ling Gomiero, Tiziano King, Ronnel Leung, Beeto Lin, You Zhen Penelo, Eva Rohe, Anna Sundseth, Øyvind Wu, Chiao Ying Xu, Jian Ping Yang, Xin Sophie Ghamarani, Amir Zhang, Wenjing Lv, Shaobo
9:00									
9:30									
10:00	***Keynote Address 5*** ***Recent developments in international testing*** ***Van De Vijver, Fons*** Leong, Frederick (session chair)	***Special Session 5*** ***Informing about ISO 10667 - An International Standard for Assessment Service Delivery in Work andOrganizational Settings*** Born, Marise Bartram, Dave Nielsen, Sverre Geisinger, Kurt Tong, Alex Harris, William G.	***S24*** ***Wechsler, Solange*** Oakland, Thomas Hutz, Claudio Byrne, Barbara	***S25*** ***Chan, Agnes Sui-Yin*** Cheung, Mei-Chun Sze, Sophia, Lai-Man Chang, Sonia	***Invited Symposium 10*** ***Zhang, Jianxin*** Zhang, Minqiang Gan, Yiqun Huang, Zheng Wang, Li Zhang, Houcan	***Invited Symposium 11*** ***Hambleton, Ronald K.*** Wandall, Jakob Hamp-Lyons, Liz Hattie, John Ercikan, Kadriye von Davier, Alina A.		***C20*** ***Guo, Fanmin*** Ackermann, Kirsten Megherbi, Hakima Mulhern, Gerry Lee, Tony Ruiz-Primo, Maria Araceli Thomas Ahluwalia, Nancy	
10:30									
11:00									
11:30-11:45	Tea Break								
11:45-13:00	Closing Ceremony and Plenary Session (LT1) State-of-the-art Speech by John Hattie								

MORNING WORKSHOPS
18 July 2010 (Sunday) 9:00am – 12:30pm

Workshop 1 Room 201



Introduction to structural equation modeling

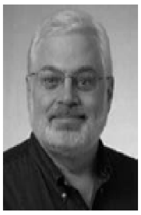
Chan, Wai (Chinese University of Hong Kong, Hong Kong SAR, China)

Abstract

Structural equation modeling (SEM) is one of the most widely used statistical techniques in social and behavioral sciences. The purpose of this workshop is to provide participants with a basic introduction to the concepts, practices, and applications in SEM. In particular, we will cover topics including symbols and notations used in SEM, path analysis, confirmatory factor analysis, and models with latent variables. The Bentler-Weeks model will be explained and described in terms of its connection with EQS, a statistical software program in SEM. In this workshop, we will minimize our emphasis on technical or statistical details of SEM. It would, however, be helpful if participants have had some training and experience with linear regression analysis.

Interested participants may consider registering for another advanced SEM session which follows this introductory workshop, where participants will have an opportunity to learn how to apply SEM using EQS.

Workshop 2 LT5



Evaluating test quality as users and writing manuals as authors: Two sides of a coin

Geisinger, Kurt F. (Buros Center for Testing, University of Nebraska-Lincoln, USA)

Abstract

This workshop is aimed at two audiences: test developers and test users. Test developers need to provide potential users with specific information so that these individuals can decide if the characteristics of the test meet their needs. Similarly, test users must look for answers to specific questions when deciding on the tests to use. These questions include intended uses of the measure, test development procedures including fairness procedures and analyses, reliability and validity evidence, the availability of norms and other scoring concerns, whether multiple forms have been developed and equated, the skills needed for test administration and score interpretation, whether the test is available in different languages (and how such new forms were developed), and whether it is appropriate for individuals with disabilities. What information needs to be made available and what may be left confidential, and where to find this information will also be discussed.

Workshop 3 LT4

Methods and designs for enhancing cross-cultural invariance

Leung, Kwok (City University of Hong Kong, Hong Kong SAR, China)

Abstract

Cross-cultural studies are regarded as quasi-experimental research, and threats that jeopardize the validity of cross-cultural differences and their explanations are reviewed. The consilience approach is advocated for strengthening cross-cultural invariance, which calls for diverse evidence based on a sound theoretical basis, multiple sources of data, different research methods, and explicit refutation of alternative interpretations. Three broad strategies for strengthening cross-cultural invariance are proposed under the consilience framework, including the systematic contrast of cultural groups, the inclusion of covariates to rule out alternative explanations, and the use of multiple research methods.

Workshop 4 LT6

Psychometric methods for investigating differential item functioning (DIF) and test bias: Concepts, methods and applications

Zumbo, Bruno D. (University of British Columbia, Canada)

Abstract

Methods for detecting differential item functioning (DIF) and scale (or construct) equivalence typically are used in developing new measures, adapting existing measures, or validating test score inferences. DIF methods allow the judgment of whether items (and ultimately the test they constitute) function in the same manner for various groups of examinees, essentially flagging problematic items or tasks. In broad terms, this is a matter of measurement invariance; that is, does the test perform in the same manner for each group of examinees? You will be introduced to a variety of DIF methods, some developed by the presenter, for investigating item-level and scale-level (i.e., test-level) measurement invariance. The objective is to impart psychometric knowledge that will help enhance the fairness and equity of the inferences made from tests. Topics include: (a) What is measurement invariance, DIF, and scale-level invariance? (b) Construct versus item or scale equivalence (c) Description of DIF methods (d) Description of scale-level invariance, (e) Examples, and (f) Recommendations.

AFTERNOON WORKSHOPS
18 July 2010 (Sunday) 1:30am – 5:00pm

Workshop 5 LT5

Establishing the ITC Guidelines on quality control in scoring, analysis and reporting of test scores

Allalouf, Avi (National Institute for Testing and Evaluation, Israel)

Abstract

In scoring, test analysis and the reporting of test scores, accuracy is essential. An inaccurate score resulting from wrong judgment, incorrect conversion of raw scores to standard scores, or accidental reporting of scores to the wrong client, are all examples of mistakes that should not occur.

Since 2008, the ITC has been developing a new set of Quality Control (QC) Guidelines for all types of measurement – psychological, educational and occupational. A draft of the Guidelines, constructed by Avi Allalouf with the help of Marise Born was distributed to eleven experts from various disciplines and different countries. They did tremendous work and their comments were of great value in revising the draft.

In the workshop the current version of the QC Guidelines will be presented and explained. Then, errors that might occur at each stage will be discussed, as well as examples and QC procedures for avoiding, detecting or correcting these mistakes. Models that deal with the causes of human error and ways to predict and reduce error will also be presented. Participants will be given hands-on practice in detecting various types of errors.

Workshop 7 Room 201

Testing basic structural equation models: Overview and hands-on application using the EQS approach

Byrne, Barbara (University of Ottawa, Canada)

Abstract

This workshop details the many stages of structural equation modeling (SEM) analyses and provides for hands-on application based on the EQS program (PC version). Following an overview of program notation and review of procedures involved in testing for the validity of hypothesized SEM models, participants are "walked through" both the specification of and results derived from the testing of two confirmatory factor analytic (CFA) and one full SEM model. Although data and software will be provided, workshop participants are required to bring their own laptops. A basic understanding of both factor analysis and SEM is a necessary prerequisite.

Workshop 8 LT6

Item Response Theory: Introduction to concepts, models, parameter estimation and fit, and several applications

Hambleton, Ronald K. (University of Massachusetts at Amherst, USA)

Abstract

Many testing agencies and researchers would like to use item response theory (IRT) models for developing, scoring, identifying bias, and equating of their aptitude, achievement, and personality tests. These IRT models, too, can be used to provide the measurement underpinnings for new test designs such as multi-stage testing and computer-adaptive testing. In this workshop, we will survey the following topics and provide several examples and practical experiences:

- Shortcomings of classical test theory that have inspired the development of IRT models, and basic classical test theoretic concepts such as reliability and item analysis,
- Specific IRT models for fitting binary and polytomously-scored data (e.g., 1-, 2-, and 3-parameter logistic models, graded response model),
- Basics of item and ability parameter estimation,
- Graphical and statistical approaches for assessing model fit (e.g., RESID PLOTS-2),
- Introduction to IRT software (e.g., BILOG-MG, PARSCALE),
- Development of tests using item and test and target information functions, and relative efficiency,
- Computer-based testing: Issues, designs, item exposure, and advantages and disadvantages,
- Identification of potentially biased test items due to culture, content, translation, and other factors,
- Follow-up readings and research.

State-of the-Art Lecture

21st July, 2010, 11:45–13:00, LT1

Global testing, global opportunities, global challenges, and a global future for assessment

Hattie, John (University of Auckland, New Zealand)

Abstract

This session will provide a retrospect on contributions from this conference, promote major issues confronting the world of testing and measurement, provide some challenges to be confronted, and suggest an agenda for this future.

19th July, 2010, 09:15–10:45, LT1

Testing for travelers: Past and future

Roe, Robert A. (Maastricht University, The Netherlands)

Abstract

Unlike testing in other fields of science, psychological testing is essentially comparative. The prevailing technology of testing is based on the paradigm of "individual differences", which assumes that people are similar except for attributes singled out for comparison. The comparative approach to testing has worked well in homogeneous and stable communities where people spoke the same language, had the same social background and shared the same culture, that is, in well bounded regions with low social and geographic mobility. But in a globalizing world, where people continually travel and communities are poorly bounded, heterogeneous and changing, it seems to be less effective. There are two main problems: (1) the comparisons produced by tests are ambiguous as the scores reflect other sources of variation (e.g. demographics, culture, language, and time); (2) competing instruments for testing travelers may give different results and it is unclear which test (e.g. which publisher, country, language and date of creation) can best be used. How can these problems be resolved? This keynote argues that psychometrics as we currently know it is unlikely to provide effective solutions. Therefore, it proposes a change in perspective that might lead to another way of testing. Starting from a historical look at how psychological testing has developed in a global environment characterized by diversity and inequality, it highlights the parties involved in testing. It claims that recognition of multiple interacting actors, with their diverging roles, views, and interests, may on the one hand reveal conflict but on the other provide a basis for developing a novel paradigm in which tests get a new purpose and format better suited for the global world of travelers.

Keynote Address

20th July, 2010, 11:00-12:00, LT1

From indigenous to cross-cultural personality assessment: the usefulness of the combined emic-etic approach

Cheung, Fanny M. (Chinese University of Hong Kong, Hong Kong SAR, China)

Abstract

Learning from the experience of adapting imported measures and following international guidelines in test development, cross-cultural psychologists have developed indigenous instruments that capture important dimensions of personality for the local cultural contexts. In this address, I will illustrate the combined emic-etic approach using the program of research involved in developing the Chinese Personality Assessment Inventory (CPAI). Incorporating indigenous and universal dimensions provides an opportunity to explore the universal and culture-specific dimensions of personality. These dynamic exchanges in cross-cultural personality assessment confront the challenges of "intellectual imperialism" in adopting translated measures and re-examine the controversy of the universality versus cultural specificity of personality structure.

20th July, 2010, 09:00-10:00, LT1

International high-stakes online testing: best practices for test security and data privacy

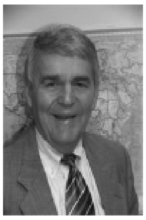
Foster, David (Kryterion, Inc.)

Abstract

More than ever it is important for organizations to measure the skills, talents and knowledge of people worldwide. Many certification programs, such as those in the information technology, medical and financial industries, are global in scope. University and college admissions programs receive applications from around the world. Pre-employment screening exams are used by companies to recruit potential employees at a worldwide level. Online universities and colleges are offering need to evaluate their students' knowledge regardless of where they live. The mobility of the worldwide workforce and student population, and the use of the Internet for marketing, communication, education, and assessment are recent and important factors which support these trends. As these tests and assessments lead to important or high-stakes decisions that affect the lives of individuals it is important that they are psychometrically sound, administered securely, and protect the privacy of examinees. The latter two are the subject of this keynote address. It will address several important questions. What are the more critical security risks when testing globally? What can be done today to reduce these risks? What promising new innovations in security are on the horizon? What are the specific data privacy issues to consider when providing a global testing program and enforcing critical security rules.

19th July, 2010, 14:00-15:00, LT1

Ethical and other professional issues: what to do when working in the absence of local standards



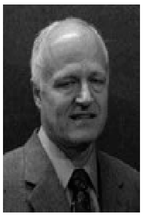
Oakland, Thomas (University of Florida, USA)

Abstract

Psychologists and other professionals recognize the need to know and adhere to the ethics code in the country or countries in which they work. However, most countries do not have ethics codes that govern the work of psychologists. Thus, psychologists working in countries that do not have an ethics code face a dilemma: they need to behave ethically yet do not know the guidelines or standards that govern these behaviors. Some cross-national conditions about which psychologists should be aware when working cross-nationally, especially in countries that may lack an ethics code, are discussed. These include knowledge of the host country's prevailing moral values, its laws and administrative policies, and ethics codes as well as policies approved by international agencies and associations. Eight guidelines applicable to test use are provided for psychologists working in host countries that lack ethics codes.

19th July, 2010, 11:00-12:00, LT1

Validation support for selection procedures



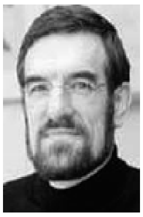
Schmitt Neal (Michigan State University, USA)

Abstract

Various means of supporting the use of selection procedures will be briefly summarized. The preponderance of the evidence on validity comes from criterion-related validation studies conducted over the last century primarily in the U.S. and Europe and very usefully summarized in meta-analyses. These meta-analyses indicate that measures of many constructs have practically meaningful implications for organizations and individuals. We also discuss the limitations of the primary data base that is the basis of these meta-analyses and propose a collaborative longitudinal data collection effort that would involve multiple organizations in various countries to address these limitations.

21st July, 2010, 10:00–11:00, LT1

Recent developments in international testing



van de Vijver, Fons J. R. (Tilburg University, the Netherlands and North-West University, South Africa)

Abstract

Cross-cultural aspects of psychological tests are increasingly important; for example, tests are adapted for use in new settings, assessment is conducted in multicultural groups, and recruitment is more and more done from an international applicant pool. The presentation will describe recent developments in cross-cultural assessment that are relevant for such applications. The topics are:

- beyond emic and etic measurement: toward a balanced treatment of culture in assessment;
- recent developments in test translations and adaptations;
- cultural loadings in assessment;
- recruiting for a global economy;
- bias and invariance testing.

Plenary Session Speaker

19th July, 2010, 09:15–10:45, LT1

Equity Interest in Testing and Measurement from an Interested Observer

Wilson, David A. (Graduate Management Admission Council, USA)

Abstract

There are those today who would argue that, indeed, the world is not improved by the measurement of competence or by psychological testing. They resist the accountability that is concomitant with measurement and decry the establishment of standards of performance.

Others demean the science of psychometrics and psychological testing through the creation and aggressive positioning of ill-conceived or poorly designed assessments.

And yet, in the face of these assaults on the profession, there seems to be little initiative on the part of the leadership to take up the charge; to speak with a strong and public voice about the need for standards and about the dangers inherent in incompetent measurement. If we lack the courage and resources to take up this challenge, we run the risk of indeed being irrelevant.

Special Sessions

Monday, 19th July 2010

Special session 1 11:00-12:00 LT5

Debating the cost of psychological tests and the factors that determine the cost

Foxcroft, Cheryl (Nelson Mandela Metropolitan University, South Africa)
Bartram, Dave (SHL Group Ltd, United Kingdom)
Foster, David (Kryterion, Inc., USA)
Oakland, Tom (Department of Educational Psychology, University of Florida, USA)

In many countries, test users and practitioners have raised concerns about the price of psychological and educational tests. Especially in countries with emerging economies the high cost of tests impacts negatively on good assessment practices (e.g., pirating and test users not purchasing the original versions from publishers). A balance needs to be achieved between raising sufficient funds for the further research and development of tests through test sales while also ensuring that practitioners who need to use the tests can in fact access them. This panel discussion will explore the factors related to determining the price of tests, and debate whether we are achieving the balance between getting enough revenue from test sales to research and develop tests but still keeping the price affordable for test users.

Special session 2 14:00-15:00 LT6

International Test Commission Guidelines for adapting educational and psychological tests (2nd edition)

Bartram, David (SHL Group, England)
Gregoire, Jacques (University of Louvain, Belgium)
Hambleton, Ronald (University of Massachusetts, Amherst, USA)
van de Vijver, Fons (University of Tilburg, the Netherlands)

Interest has been growing for years in the topic of translating and adapting educational and psychological tests from one language and culture to others. Today, many of the popular intelligence and personality tests are translated and adapted into 50 or more languages; achievement tests such as those used in the large scale international assessments like TIMSS and PISA are translated and adapted into more than 30 languages and cultures. In the United States, as one approach for handling cultural diversity and accommodations, many states are making their achievement tests available to students in more than a single language. Tremendous progress too has been made in the methodology for translating and adapting tests. The journals are publishing many articles on topics such as conducting judgmental reviews, studying construct equivalence, and identifying item level bias. The purpose of this session is to introduce the second edition of the International Test Commission (ITC) Test Adaptation Guidelines. The participants will discuss the 17 new Guidelines, the process used to develop them over the past three years, comparisons between the first and second edition of the Guidelines, and their relevance for test adaptation practices in cross-cultural assessments. The four participants in this presentation are part of the six person ITC committee who had the responsibility for producing the second edition of the Guidelines.

Tuesday, 20th July 2010

Special session 3 — 09:00-10:00 LT2

The International Journal of Testing: Ten years and going strong

Hattie, John (University of Auckland, New Zealand)

The status and purposes of the Journal will be outlined, and the notion of "international" and "testing" outlined. There will be a presentation on behalf of the current editors (Steve Sireci and Rob Meijer) and previous editors will talk about their experiences and answer questions.

Special session 4 — 11:00-12:00 LT6

Examining formative assessment

Bennett, Randy (Educational Testing Service)
Brown, Gavin (The Hong Kong Institute of Education, Hong Kong)
Koh, Kim (Nanyang Technological University, Singapore)

In elementary and secondary education, formative assessment is in vogue. A key reason for the popularity of formative assessment is, undoubtedly, the claims that have been made for its effectiveness. This session reviews the evidentiary sources cited for these claims, and summarizes how the effectiveness of formative assessment might be responsibly represented. Two discussants will react, including in their comments the results of their own research on teachers' beliefs and their formative assessment practices, and on the potential emotional and social consequences students may experience through self- and peer-assessment activities.

Wednesday, 21st July 2010

Special session 5 — 10:00-11:00 LT2

Informing about ISO 10667 - An International Standard for Assessment Service Delivery in work and organizational settings

Born, Marise (Erasmus University Rotterdam, the Netherlands)
Bartram, Dave (SHL Group)
Nielsen, Sverre (The Norwegian Psychological Association, Norway)
Geisinger, Kurt (Buros Center on Testing & University of Nebraska, USA)
Tong, Alex (ATA Inc., China)
Harris, William G. (Association of Test Publishers, UAS)

In March 2007, in Berlin, the Deutsche Industrielle Norm (DIN) took the initiative to develop an International Standard for Assessment Service Delivery, known as ISO 10667. This Standard is process-oriented and focuses on procedures and methods to assess people in work and organizational settings. In the Standard, a pre-assessment phase, an assessment delivery phase and a post-assessment review phase are distinguished. International Standards are developed by a process of several formal stages. In July 2010, ISO 10667 has entered the Enquiry stage. In this stage, the draft International Standard (DIS) is commented upon by ISO member bodies. Within this session, we offer a briefing on the content of the Standard and the stage its development is in. A panel of participants in ISO 10667 will subsequently discuss the Standard. The audience will be given the opportunity to look at the draft International Standard (DIS).

Monday, 19th July 2010

Diamond sponsor session 1 11:00-12:30 LT6

More than scores

Chair

Anastasio, Ernest J. (Graduate Management Admission Council, USA)

Symposium Abstract

Test taker databases include vast amounts of valuable information in addition to test scores. Demographic data and information about where examinees send their scores, for example, can provide actionable, current insight into the global marketplace. This session will discuss ways the test sponsor can collect, organize, and disseminate information that can be gleaned from test databases using the Graduate Management Admission Test® (GMAT®) as an example. First introduced 55 years ago, more than 50 percent of all GMAT® examinees today are non-US citizens. Annual publications such as the GMAT® Candidate Profile and Geographic Trend Reports (World, Asian, and European) as well as individualized reports and validity studies will be highlighted. During this session we will discuss how each of these services were designed and implemented to more efficiently reach our increasingly global client needs.

Paper 1

"Reporting examinee population demographic changes"

Guo, Fanmin (Graduate Management Admission Council, USA)*

Changes have been occurring in the examinee population of many large-scale assessments for many years. Technological and psychometric advances have allowed the expansion of test use around the world and an increasingly diverse demographic of examinees. This presentation discusses two topics in light of such changes. The first is how to report the demographic changes using five-year rolling data for profiling. In addition to a printed copy of the profile report which is distributed to GMAT using schools every year, an interactive web-based version is provided to make more detail available to Graduate Management Admission Council (GMAC®) member schools. Details of both forms of this report will be presented. The second topic describes the redesign of the printed five-year profiles to switch focus from the needs of US schools to that of schools worldwide. Some demographic variables were redefined with an international perspective and others have been removed or reformatted in the profile to fit our new global client base. The process used to identify the changes needed to be more responsive to the global marketplace, including the design and implementation of a global survey, will be discussed. The presentation will be facilitated using examples from the printed and interactive profiles.

Paper 2

"Updating validity data collection and management"

Talento-Miller, Eileen(Graduate Management Admission Council, USA)*

Although test sponsors are keenly aware of the need for continuing evidence for the validity of test scores for their intended use in various populations and situations, the users themselves often do not feel as pressured to study their own use of scores, trusting in the quality of the test based on previous research or non-statistical evidence. In the case of admission tests, helping schools understand the validity of scores for their specific current programs benefits not only the users but the test sponsors as well, expanding the evidence available and keeping up to date with changes that may stem from changing demographics in the examinee population. Continuing validity evidence is necessary to ensure the efficacy of scores for admission to global business programs. This presentation will elucidate how the sponsors of the Graduate Management Admission Test (GMAT) exam have updated the Validity Study Service to streamline data collection and management and improve the quality of the reports to ensure value for the users. In addition, the presentation will discuss the methodology behind various meta-analyses that have been conducted to summarize the data from diverse programs around the world.

Paper 3

"Increasing options: Communicating shifts in interest in international programs"

Defibaugh, Courtney (Graduate Management Admission Council, USA)*

As with many testing programs, the population of Graduate Management Admission Test (GMAT) examinees has been evolving over the years. As such, demands from the users of the GMAT have increasingly shown an interest in test takers outside of the United States asking such questions as "Who are they?" "Where do they want to study?" To reach this demand a new set of reports, Geographic Trend Reports, was developed by Graduate Management Admission Council (GMAC). The Geographic Trends series currently consists of four reports; World, Asian, European, and North American. The World Geographic Trend Report reveals score sending patterns for citizens of 10 regions of the world. Drilling down farther, the Asian and European Trend Report provides data analyses on the top 10 citizenship groups for the Asian and European World Regions. The North American Trend Report shows trends on those tests taken within the United States and Canada. This session will cover how the Geographic Trend Report series was designed and implemented to meet the needs of GMAC clients.

Paper 4

"Handling information requests: How and why"

Taliaferro, Hillary (Graduate Management Admission Council, USA)*

With background questions and historical data linked to test scores, testing companies have a resource that, if properly utilized, can be enormously helpful to test users. While the Graduate Management Admission Council (GMAC) produces a wide range of publically available publications, some school users desire more detailed and specific data analyses to understand their own marketplace—who the examinees are and where they are interested in studying—and help them to develop marketing strategies. Through these special information requests, data can be aggregated to focus information to better meet the needs of our global clients. A simple information request may require only mean total scores for a small group of examinees, such as students near a particular city. Responses to more complex requests may consist of numerous comparative graphs and tables. This session will highlight the types of questions and the processes used to complete both simple and complex information requests received from our clients around the world. Examples will show end products used in presentations, research projects, and marketing. Going beyond basic scores and standard publications helps build relationships and underscores that an organization is truly client oriented.

Discussant

Rudner, Lawrence (Graduate Management Admission Council)*

Discussant will summarize symposium session. The discussant will describe possible applications at an international level for products described during the session.

Tuesday, 20th July 2010

Diamond sponsor session 2 13:30-15:00 LT6

Testing across cultures

Chair

Schuchart, Nadine (Hogrefe Verlag GmbH & Co. KG, Germany)

Symposium Abstract

The drive to improve methodology for achieving cross-cultural equivalence when adapting psychometric tests has never been more important. In a world of increasing cultural diversity both across and within nations, the task of ensuring sensible interpretation and fair comparisons has grown in its complexity. Add to this the emergence of new testing formats and an evolving number of constructs of interest; the achievement of a solution to culture free testing seems to be ever more like a minefield. Indeed the first paper in this symposium questions the feasibility of achieving a solution at all and warns against achieving equivalence at the expense of the construct being measured. The second paper compares the functioning of a work based personality questionnaire across a number of European cultures giving special attention to item functioning rather than restricting the study to scores at the overall scale level. The third paper considers the special issues in adapting the same work based questionnaire to China; a culture very different from those across Europe. Our final paper presents findings from the cross-cultural adaptation of a Leadership Judgment questionnaire comparing the generalizability of leadership style preferences, leadership judgment and item difficulty level across a range of European countries.

Paper 1

"How to test the same thing in different cultures?"

Gillet, Isabelle (Editions Hogrefe France S.A.S., France)*

In the I/O field, the construction of scales to be deployed in multiple cultures/languages is a key issue for tests developers. ITC guidelines propose steps, in order to guarantee fairness, equity and equivalence across national versions. Tests users demand that different language versions have the same scales, the same number of items and the same items thinking this proves the equivalence of testing across countries; as if "people are all the same, all around the world,." This paper addresses some specific questions regarding the dangers of "smoothing" away real difference to achieve surface equivalence such as:

- "Missing" the country specific aspects of how constructs manifest in behavioural terms
- Throwing out genuine variance when we get rid of a scale which is not strictly equivalent in each country?
- Dropping items which prove to be very good in one country but not in another one?

The fact is that tests assess human beings; that human beings express themselves through language, values and behaviors (all aspects greatly impacted by culture); so adaptation of tests must therefore take account of this important reality... culture makes a difference.

Paper 2

"Measurement and structural equivalence of European versions of a work-based personality questionnaire"

Li, Tao (Hogrefe Ltd., UK)*

Measurement equivalence is an indispensable requirement for valid cross-cultural comparisons. This paper explores the adaptation of "The Business-focused Inventory of Personality" (BIP) to a range of European cultures. The BIP has been selected as the focus for the paper because it was originally developed in Germany. It is somewhat unusual to have the source language for test adaptation being anything other than English (most commonly US English but more recently British English too). The comparative analysis across the different national versions has, in this study, been carried out at an item level as well as the usual mean score comparison level. This is considered to be important since differential item functioning may be masked when only overall scale scores are compared; thus important information for cross cultural comparison can be obscured. The paper demonstrates the application of multiple indicators, multiple causes (MIMIC) structural equation model, a relatively new technique, in detecting differential item function.

Paper 3

"Adapting a European work based personality test to China"

Benoit, Andreas (Benoit Consulting, Hong Kong SAR, China)
Roth, Hans J. (Swiss Consul General in Hong Kong, Hong Kong SAR, China)*

In an ever more interconnected world, with economic aspects of life increasingly linked to socio-economic development, organizations will need to intensify the development of talent in the workforce in order to keep pace with the increasingly intense competition. In order to measure and further develop leadership and other stylistic competences, there is a need for top level diagnostic tools specifically adapted for the Chinese market to identify current individual strengths and weaknesses. The BIP is a work focussed personality test developed in Europe. BIP – "made in China" – will be tailored to the specific culture and norm-groups of China and will serve the particularly strong and growing need for talent development in this part of the world. This paper will focus on:

- Cultural differences and peculiarities which need to be considered when adapting the tool
- Procedures for translating into Mandarin and for data collection
- Presentation of the first results and feedback from Chinese users and professionals

Paper 4

"The development of a situational judgment test from a cross-cultural perspective"

Li, Tao (Hogrefe Ltd., UK)*

Situational Judgement Tests (SJTs) are increasingly popular but there are less frequent reports on the specific considerations arising from such tests regarding international adaptation. This paper will focus on what has been learned from adapting a British SJT, 'the Leadership Judgement Indicator' (LJI) to five European countries. The LJI is of particular interest in terms of the adaptation process, because in addition to measuring how effectively leaders judge the style most appropriate for the situation, there is also a measure of preferred style. This allows a comparison of the cross-cultural generalisability of each type of construct when measured in an SJT format. The LJI items are scored for judgement according to goodness of fit to a theoretical model such that the difficulty level of items is a further comparison of interest cross-culturally. Finally, given the model on which the LJI is based arises from Western leadership theory a question is raised as to the impact of this on generalisability of the test to Eastern cultures.

Discussant

Schuchart, Nadine (Hogrefe Verlag GmbH & Co. KG, Germany)*

Monday, 19th July 2010

Invited symposium 1 11:00-12:30 LT2

Exhibition on testing & measurement

Organizer and Moderator

Allalouf, Avi (National Institute for Testing and Evaluation, Israel)

Participants

Allalouf, Avi (National Institute for Testing and Evaluation, Israel)
Bennett, Randy (Educational Testing Service, USA)
Hambleton, Ronald (University of Massachusetts, USA)
Zhang, Houcan (Beijing Normal University, China)
Grégoire, Jacques (Université Catholique de Louvain, Belgium)
Gafni, Naomi (National Institute for Testing and Evaluation, Israel)

Abstract

A scientific exhibition on testing & measurement is currently being developed. NITE (National Institute for Testing & Evaluation) and the Bloomfield Science Museum, Jerusalem have already begun working on the scientific exhibits. The current partners are ETS (Educational Testing Service) and the Franklin Institute, Philadelphia. Among the topics presented by the exhibition will be: the history of testing and its role in society, reliability, validity, intelligence measurement, psycho-physiological measurement, psychological assessment, selection and vocational assessment, international comparisons, gender impact, test preparation and coaching, cultural aspects of testing, fairness and bias, adaptive testing, technology and the future of testing. In addition, the exhibition will deal with the challenge: how do we measure characteristics that cannot be measured directly? The exhibition includes some 25 exhibits, most of them interactive. Some of the activities are designed for groups. In addition, the exhibition will include posters, videos, photographs and old test forms. An internet website will be created which can be accessed before and after the exhibition. The exhibition team is already working on the exhibits. It consists of experts in several fields: psychometrics, science, museology and internet design. A public opinion questionnaire has been formulated and data is being gathered and analyzed. We believe that this session will benefit the ITC audience and would be an important addition to the conference program. We conceive the exhibition on testing & measurement as an innovative means of familiarizing the public, youth in particular, with educational and psychological measurement concepts. Dissemination of measurement concepts is one of the organization's main goals. The session offers an open and fruitful discussion with distinguished measurement experts on how to reach the exhibition goals, see below.

Invited symposium 2 15:45-17:15 LT5

The use of ethical principles in testing, - or the lack of it

Chair

Nielsen, Sverre Leonard (The Norwegian Psychological Association, Norway)

Symposium Abstract

This symposium is an attempt to focus on ethical issues in assessment and testing. While quality assurance of tests and testing in general gains more and more attention, the explicit focus on ethics both in regulations and daily use are not so easy to see. Quality assurance of both competence and methods is in itself connected to ethical principles. However, there is still a challenge of raising the ethical awareness among test users.

Paper 1

"Testing ethically: Tensions between principle led ethics and regulatory system"

Lindsay, Geoff (Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, UK)*

There has been a large scale interest in the development of ethics by psychologists across the world. The number of countries with an ethical code has increased, stimulated and supported by international initiatives. Within Europe, The European Federation of Psychologists Associations approved a Meta-code of Ethics in 1995, revised in 2005: all member associations are required to have an ethical code compliant and not in conflict with the Meta-code. At a world-wide level, three international associations of psychology (IUPsS, IAAP and IACCP) approved a Universal Declaration of Ethical Principles for Psychologists in 2008 following collaborative, developmental work by an ad hoc group comprising senior psychologists from across the world, a strategy to optimise sensitivity to cultural issues. There have also been capacity building initiatives to support psychological associations in the early stages of development of their code, most recently in South East Europe. Common to many codes developed by national associations is a focus on the use of ethical principles and in some cases supportive material and training in ethical decision making, using the principles and the national code's specifications/standards of behaviour. However, there has also been interest in many countries to attain statutory regulation of the profession, in which case consideration of allegations of unethical conduct may be heard by a separate statutory body with its own code of conduct. In this paper I shall explore the tensions that can arise from the application of each of these two ethical systems

Paper 2

"Ethical guidance when local standards do not exist"

Oakland, Tom (University of Florida, USA)*

Data from two international surveys of test development and use with children and youth, one conducted in 1990 and the other conducted within the last few months, reveal a number of important changes, some of which have important implications for test ethics. These changes are summarized and implications regarding test ethics are described.

Paper 3

"When is assessment a 'Psychological Act'?"

Bartram, Dave (SHL Group Ltd, UK)*

The issue I address is that of deciding when an assessment or some part of an assessment becomes a 'psychological act' and whether 'psychological acts' necessarily require the intervention or input of a practicing psychologist. In considering this issue, we need to consider assessment processes as involving a number of stages and consider how the need for specialist expertise may apply to each stage and under what conditions. I will also consider the degree to which such acts might be carried out remotely or by proxy, and under what conditions. This is an issue that sits at the heart of much of the discussion over the use of the Internet in assessment, especially the issue of remote administration. This has been a topic of much debate over the past few years culminating in the publication of Tippen's (2009) focal paper and accompanying commentaries. I conclude by arguing that ethical assessment depends on the competence of assessors. Professional labels do not guarantee competence and hence we need to relate the nature of the acts carried out in assessment to the competence required by the actors.

Paper 4

"The lack of focus on ethical issues among counsultant within the IO-field"

Hagås, Per Olav (Manpower Professional Executive AS, Norway)*

The issue of ethics is of great importance and affects headhunters, recruitment companies and agencies which let out personnel. Most headhunters and recruiters have a background from sales and management. Few headhunters are psychologists, nor do most of them have adequate competence within objective assessment of human character and/or tools which measure human potentials. Personality tests in particular, but also ability tests are frequently and increasingly used in connection with employments in all categories on any level. The presentation will address the use of test tools and what demands are set to the tests that will affect the careers of the tested ones. Further, what demands that will be profitable expect of the tests and the ones using them? Finally, I will point out the ITC's International guidelines for test use, and propose which parts of these guidelines that should be set as requirements to serious commercial traders. The conclusion is that qualitative demands and ethical norms to a lesser extent have adequate focus, and that this is an issue which is extremely interesting to address.

Invited symposium 3 15:45-17:15 LT6

How to apply personality as a worldwide common concept: Big Five, Big More, or should we? Modeling and usefulness.

Chair

Weekers, Anke (Cito, Netherlands)

Symposium Abstract

The 'Big Five' personality concept is often put forward as a unifying comprehensive framework to describe adult personality worldwide. At the same time, debate goes on regarding four main issues. First, are 'Five' factors enough or are more factors needed to account for the empirical data collected so far. Second, do persons respond to personality items in the way their responses are modeled. More specifically, which factors might give rise to different response strategies. Third, is personality a useful concept in the prediction of organizational behavior for selection purposes, both regarding the incremental validity above other factors and regarding ethical issues. Fourth, how cross culturally valid is any personality framework. The present symposium will discuss the mentioned issues against the background of globalization of test use. Questions like the following then become relevant. Which personality factors are cross culturally valid and which might be specific for which culture? Are there cultural specifics as regards disclosing information on personal behavior? In particular, to what extent is the self report methodology as in personality questionnaires equally applicable in various cultures and which models have to be used to model response processes over various cultures? Are there cultural specifics as regards the predictive validity and ethics of test use in applied psychological practice? Contributors to the symposium will focus on one or more of the above aspects. In a general discussion similarities and differences between cultures will be discussed and consequences will be formulated for both research and practice.

Paper 1

"Impact of measurement invariance on construct correlations, mean differences, and relationships with external correlates: Big Five and RIASEC Measures"

Schmitt, Neal (Michigan State University, USA)*

A relatively large number of cross-cultural studies have investigated the invariance of measures used with various groups and a common finding is that statistically significant differences between groups do exist. In this paper, we evaluate the importance of this lack of invariance on the estimation of structural parameters that relate constructs in these studies. Specifically, the impact of measurement invariance and the provision for partial invariance in confirmatory factor analytic models on factor intercorrelations, latent mean differences, and estimates of relationships with external variables is investigated for measures of two sets of widely assessed constructs: Big Five personality and the six Holland (1985) interests (RIASEC). In comparing models that include provisions for partial invariance with models that do not, the results indicate quite small differences in parameter estimates involving the relationships between factors, one relatively large standardized mean difference in factors between the subgroups compared, and relatively small differences in the regression coefficients when the factors are used to predict external variables. The results provide support for the use of partially invariant models, but there does not seem to be a great deal of difference between structural coefficients when the measurement model does not include separate estimates of subgroup parameters not invariant. Future research should include simulations in which the impact of various factors related to invariance is estimated.

Paper 2

"The Big Five in selection and development assessment: 'Blessing and yoke'

De Fruyt, Filip (Ghent University, Belgium)*

The Big Five has generated a flourishing stream of research leading to a revitalized interest during the past fifteen years in personality assessment in Industrial, Work and Organizational (IWO) psychology. Although this enthusiasm has been welcomed by many, it also has been the subject of considerable criticism (see the debate between Morgeson et al., 2007 versus Ones, Dilchert, Viswesvaran and Judge, 2007; Tett and Christiansen, 2007). Rather than reiterating these arguments, the current contribution highlights new avenues to better align personality research and IWO applications, taking benefit from recent advancements in personality assessment including 'trait-activation theory', 'frame-of- reference research', 'person-centered approaches' and 'a spectrum conceptualization of traits'. Their impact on applied personality assessment will be illustrated using data collected with the Personality for Professionals Inventory (Rolland & De Fruyt, 2009) administered to samples of students and incumbents.

Paper 3

"Response processes in personality measurement"

Weekers, Anke (Cito, Netherlands)*

In the employment context most self-report personality inventories are constructed using classical test theory, factor analysis or dominance IRT models. These models assume dominance response processes, and were originally developed for maximum performance measures (what someone can do), like ability measures. Although personality traits are typical performance constructs (what someone usually does), personality inventories are developed according to the same assumptions as used in maximum performance measurement. However, persons might respond differently to self-report personality inventories, and a different kind of response processes, the unfolding or single-peaked response processes, might be more likely. The usefulness of these response processes will be discussed. An example will be given of an inventory measuring Order that is developed based on single-peaked response processes. The original inventory was developed in the USA (Chernyshenko, Stark, Drasgow, & Roberts, 2007), and translated in Dutch. For this research the translated scale is used, but results will be compared to the results found in the USA.

Paper 4

"Are there Big Cultural differences in Personality?"

Serlie, Alec (Erasmus University Rotterdam / GITP, Netherlands)*
Hiemstra, Annemarie (Erasmus University Rotterdam / GITP, Netherlands)
Van Leeuwen, Rob (GITP, Netherlands)
Bazen, Madelijn (Leiden University, Netherlands)

The results on personality questionnaires play a significant role in selection assessments. Predictive validity of the Big 5 in relation to job performance has been well documented both in America as well as Europe. As personality questionnaires are generally language based, there is a distinct possibility that members of minority groups might either not understand items correctly or misinterpret the meaning of the items. This may in turn bias the results of a questionnaire. Several models of test fairness have been proposed. The most widely accepted is Cleary's model, which is based regression lines. More recently models using Differential Item Functioning (DIF) have been introduced, whereby individual items can be studied in relation to the trait associated with the item. In the present study we set out to study the items of a Five Factor personality questionnaire using various DIF techniques. Our hypothesis was that there would be a difference between (ethnic) minority and (Dutch) majority respondents. In this study the data of a group of 280 test takers (minority: n=119, majority: n= 161) who had completed a FFM personality questionnaire were analyzed. All participant were in their final year of higher (Bachelor or Master) tuition. The results of the study showed that there were significant differences between the two groups on all five factors. On the item level only 11 (5,5%) items showed DIF, of which most could be found in the Neuroticism factor. Items on the Conscientiousness and Openness factors did not show any DIF.

Discussant

"Assessing personality. Does culture matter?"

Van De Vijver, Fons (Tilburg University, The Netherlands & North-West University (Potchefstroom Campus), South Africa)*

Following the four papers put forward at the symposium, their results will be evaluated critically with respect to the following issues: What are defensible empirical generalizations as far as the intercultural (in)variance of a common personality framework is concerned? More specifically, what are the merits of the Big Five factor model in this respect? Do they differ with respect to invariance, or are even additionally other factors necessary to account for intercultural differences? How do the innovative contributions that are put forward with respect to response scale methodology as well as alternative measurement concepts compare with classical big five questionnaire methodology? What are advantages and what might be pitfalls, especially with respect to intercultural differences in response scale use? From a prediction perspective, the incremental validity of more sophisticated scoring methods as well as that of taking situational specifics into account will be discussed. Consequences for advancing theory as well as improving practical utility will be suggested.

Invited symposium 4 13:30-15:00 LT5

The current State of Psychological and Educational Testing and Assessment in Indonesia

Chair

Purwono, Urip (Universitas Padjadjaran, Bandung, Indonesia)

Symposium Abstract

This symposium displays the recent development, current practices, and research in the use and development of psychological and educational testing in Indonesia. The expectation is that participants will have an updated perspective on the landscape of testing in the developing countries such as Indonesia where, in one hand, the number of scholars with adequate background training in psychometrics, test and measurement are limited and, on the other hand, the challenges and work to be done are abundant. The symposium is also aimed at getting thoughtful ideas from the participants. The papers presented in this symposium includes: (1) Mathematics Test Equating under the Graded Response Model; (2) Development and Preliminary Validation of Musical Ability Assessment System; (3) How is IRT applied in personality tests?: A study using Indonesian sample; (4) Improving Students Achievement, Learning Responsibility, and Learning Behavior in Mathematics using Assessment for Learning Model; (5) A closer look at the Indonesia's National Exam Program; and (6) The Uses of Tests in Psychological Practices and Research in Indonesia. Overall, the paper presented will represent (a) the common practices of making assessment instruments available in Indonesian language; (b) the use of test in the common psychological practices and research in Indonesia; (c) current advances in measurement research; and (d) attempts to make a commonly individually administered test available for a large testing program as well as the current trend of collaboration between psychometrician and practitioners as well as researchers in the area beyond psychology and education in Indonesia.

Paper 1

"Improving students achievement, learning responsibility, and learning behavior in mathematics using assessment for learning model"

Mansyur, Mansyur (The State University of Makassar, Indonesia)*

Linking assessment with learning has been a concern in educational assessment for the last two decades. While many attempts have been made to construct assessment that facilitates students' learning, experimental study investigating the effect such assessment program to different aspects of learning was limited. This study investigates whether an assessment for learning program increases (1) student achievement in mathematics, (2) student responsibility in learning mathematics, and (3) student behavior in learning mathematics. A single-group interrupted time-series design (Creswell, 1994) was employed in this study. A total of 10 meetings in-class "Assessment for learning" program was conducted to 244 grade 7 public schools students in Makassar. Assessment instruments consisting of (1) a two stages assignments to assess student's mathematical achievement, (2) self report of learning responsibility, and (3) observation list to be used by teachers were developed. Students' achievement, responsibility, and behavior in each session were estimated employing Samejima's graded response model. Differences of ten data points were analyzed. The results showed that by the end of the program 80.79%, 76.76%, and 71.19% of the students show high achievement, responsibility, and expected behavior consecutively. It is concluded that the assessment for learning program had a positive effect on students' achievement, responsibilities, and behavior in learning mathematics.

Paper 2

"How is IRT applied in personality tests: A study using Indonesian sample"

Halim, Magdalena (Catholic University of Atmajaya, Jakarta, Indonesia)*

Item Response Theory (IRT) is a modern psychometric approach, which provides valuable methods for the evaluation of psychological measurements, including objective personality assessment. Although most published applications of the IRT are used to examine cognitive and ability tests, IRT models are also increasingly being used to study psychometric characteristics of personality tests nowadays. IRT models differ from one another in the number of modeled parameters. Within the family of IRT models, the two-parameter logistic model (Birnbaum, 1968) has been chosen in the present study. The purpose of this study is to show the applicability of IRT analysis in evaluating the psychometric properties of an Indonesian version of the MMPI-2. A total of 1,473 individuals completed a valid MMPI-2; 63.1% were women and age ranged from 17 to 61 (M=23.84, SD=7.38), the rest (36.9%) were men and age ranged from 17 to 66 (M=26.38, SD= 9.29). This sample is part of the Indonesian MMPI-2 normative study. The item parameters for the two-parameter model were estimated with BILOG 3 program (Mislevy & Bock, 1990). The results of this study considerably add to existing knowledge about the psychometric properties of MMPI-2 in Indonesian sample. The two-parameter IRT model fits the data quite well and can be considered as appropriate model. The fact that these results were obtained in an Indonesian sample could be also criticized for being culture specific.

Paper 3

"Development and preliminary validation of Musical Ability Assessment System"

Salim, Djohan (Indonesia Art Institute, Jogyakarta, Indonesia)*
Purwono, Urip (Faculty of Psychology, Universitas Padjadjaran, Bandung, Indonesia)

Musical ability is an important aspect of human quality. From a neurological point of view, individual ability to perform and comprehend musically appears to work independently from other forms of intelligence. This paper reports an attempt to assess individual musical ability using a paper and pencil approach. Defining musical ability as a basic ability to recognize, distinguish, reproduce, and perceive similarity and differences between various sounds, the assessment is carried out by presenting individuals with pre-recorded sounds produced by musical instruments with variation in pitch, timbre, tempo, and dynamics. The individual tasks are to answer questions related to the characteristics of the sound. Standardized is done by pre-recording the sounds in a CD. A total of 200 participants were recruited for the study. Partial EEG activities were also recorded for validation purposes. Even though, in term of internal consistency, the estimated reliability was at the lower range, the psychometric properties of the assessment system show promising features. Preliminary validation also shows consistencies with neurological theory, suggesting that the idea underpinning the assessment can be developed further, particularly to facilitate research in the area of psychology and music.

Paper 4

"A closer look at the National Exam Program in Indonesia"

Mardhapi, Djemari (Indonesia National Education Standard, Indonesia)*

Since its conception back in early '60, Indonesia National Exam has been a concern to parents, educators, and policy makers. As a results of the controversies, political agenda, and advances in test theory, the format of Indonesia National Exam has been undergone several changes. This paper presents the historical, technical, delivery, and reporting aspects of the National Exam in Indonesia and the changes in its format from one period to another. Information is gathered from direct observation, reports, printed documents, publications and governmental official archives related to the design and implementation of the exam. Results of the analysis to the item level national exam data were also examined. It is asserted that, from psychometric perspectives, the current national exam has sound technical aspects. The IRT one parameter logistic models was appropriate for the data. Arising problem is hypothesized as more closely related to the non technical aspects of the exam, particularly its administration and the public misperception concerning the exam.

Paper 5

"The uses of tests in psychological practices and research in Indonesia"

Suhapti, Retno (Indonesian Psychological Association, Indonesia)*

The history of psychology in Indonesia is tied with the use of psychological testing for personnel selection, placement and classification. This tradition continues until recently, and this paper identifies issues around psychological testing and its uses in Indonesia. Data from this research was obtained from a survey administered to professional psychologist in Indonesia inquiring activities frequently engaged in their services and their uses of test materials. In addition, information was also gathered from three separate large group discussions taking places in three difference provinces. A focused group discussion was also conducted to revalidate the information. Results suggested the questionable practices related to the development and use of test materials which stemmed from inadequate background training in testing and measurement coupled with lack of guidelines and regulation concerning test uses in the country. Another finding suggested that the most frequently used psychological testing instruments were transported from either the US or Europe. However, proper adaptation of these instruments and empirical investigation of its equivalences with the parent language culture of the test is rarely conducted posing a validity concern to their uses in Indonesia. These findings are then discussed in the light of the history psychological education in Indonesia, particularly those related to testing and measurement. Recommendations for future intervention, particularly those related to the international testing community, are presented.

Invited symposium 5 15:15-16:45 LT1

Future directions of testing in the United States

Chair

Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

The importance of test uses continues to grow. In the USA today, students from the 3rd grade to high school are administered achievement tests, for a total of about 80 million tests per year. Add several times 80 million tests to account for the diagnostic tests that are being administered to support the assessment of student progress, and it is clear that the growth of testing in the schools has been substantial. Also, the number of admissions tests and credentialing exams continues to grow. At the same time, because of the importance of these tests, and the desire to improve test score validity, technical advances in many directions are occurring. In this symposium, the presenters will focus on three important directions in the USA: First, we will consider the impact of cognitive psychology on testing. The impact has been discussed in the measurement literature for 30 years but now that impact is being seen, and substantial amounts of research are underway. Second, the impact of technology has been increasing, and now it is likely to fundamentally change our approaches to what is measured, as well as to test design and test administration. Advances in technology will be the focus of the second presentation. Finally, nothing is more important, ultimately, than the production of valid test scores, that can be reported on meaningful score scales, and in ways that they are understood and used correctly by practitioners. Score reporting in the future will be the focus of the final presentation.

Paper 1

"Using advances in learning and cognition to improve assessment in the 21st Century"

Huff, Kristen (The College Board, USA)*

Progressive approaches to curriculum and instruction are strongly influenced by what we are learning about how students develop deep conceptual knowledge, critical thinking skills and the propensity to learn more advanced topics. As such, contemporary assessment design needs to be aligned with these more complex targets of learning that are valued in the classroom versus those typically measured on large-scale standardized tests such as factual recall and "plug and chug" procedures. Assessments of the future will, for example, evaluate students on the quality of the questions they have about a new topic rather than measure their mastery of content, and provide feedback that helps teachers move students along the stated learning path. In addition, strong validity arguments for contemporary assessments will include a rationale for how items are designed to explicitly target the knowledge and cognitive processes of interest. This presentation will provide an overview of progress in these areas (e.g., evidence-centered assessment design), as well as areas in need of more research (e.g., dynamic assessments of transfer).

Paper 2

"Impact of technology on testing practices"

Mills, Craig (American Institute of Certified Public Accountants, USA)*

Technology is transforming business and business processes at an ever accelerating rate. Concurrently, emerging economies are providing inexpensive access to labor, both skilled and unskilled. Widespread access to the internet and social media has changed how professionals, students, youth, and adults work, play, and interact with one another. Inevitably, these trends will impact measurement practice as well. We can anticipate that current batch processing for item development, test administration, and statistical analyses will be increasingly replaced by continuous process flows. These process flows will also become increasingly automated and decentralized. Additionally, traditional testing formats may well become less valid as success in education and work become more dependent on collaboration and use of disparate data resources. Tests will need to change in order to assess entry-level skills or educational promise as the skills needed for success change. This presentation will suggest changes in testing processes and tests themselves that are responsive to the changing needs of test sponsors and test takers. It will draw from lessons learned in other industries and apply them to test development, administration, and analysis.

Paper 3

"Next generation of Latent Variable Models: New opportunities and challenges for testing"

Zumbo, Bruno D. (University of British Columbia, Canada)*

In recent years, many exciting developments have taken place in latent variable modeling, but perhaps none more so than the development of methods that (a) explore and account for unobserved population heterogeneity, or mixtures of unobserved groups, (b) mixtures of continuous and categorical latent variables, (c) multilevel models for complex measurement data, and (d) approaches for dealing with categorical item response data. This next generation of latent variable models shows tremendous potential for improving testing practice and research by expanding the types of modeling being done and hence aligning more closely with the complex nature of contemporary tests and testing contexts. I will present an overview of this next generation of advanced methodology and describe how it may inform testing from day-to-day practices such as dimensionality assessment to more theoretical work that focuses on evaluating explanatory models of test data for the purposes of supporting validity research and practice.

Paper 4

"Improving the score scales and reports we use in communicating test results"

Hambleton, Ronald K. (University of Massachusetts, USA)*

Testing practices in education and psychology have advanced considerably in recent years through the introduction of item response theory models, generalizability theory, automated test assembly, and new test designs such as computer-adaptive testing. At the same time, methods for reporting test scores and diagnostic information to candidates, the culmination of the testing process, remain largely understudied and undervalued as a problem in educational and psychological assessment. This is most unfortunate too because of the large amount of evidence suggesting that candidates and other score users such as teachers, policy-makers, psychologists, and the media, are often confused by the meaning of test scores resulting in misinterpretations, and candidates are often disappointed by the limited amount of diagnostic information they receive from hours of testing. The goals of this presentation include: (1) highlighting several promising ideas (e.g., bench-marking and item mapping) for increasing the clarity and meaning of score reports, and (2) offering some steps, based on emerging research, for preparing score reports and score scales. Examples of good practices will be presented and come from our work with several national and state testing programs, and credentialing agencies in the United States.

Invited symposium 6 15:15-16:45 LT5

Recent developments in the assessment of emotional intelligence

Chair

Fontaine, Johnny (Ghent University, Belgium)

Symposium Abstract

Since its inception the construct of emotional intelligence (EI) has attracted a lot of attention from both practice and science. While being very popular in practice, the construct has been severely criticized in the scientific arena. The EI trait approach, which uses self-reports, is reproached to lack discriminant validity. The EI ability approach, which uses maximum performance tests, is criticized for the content of its items and for its scoring. In this symposium four recent developments that take these criticisms to heart are being presented. In the first contribution a set of four new assessment instruments is proposed. They form a multi-method approach in which EI is assessed by self-reports, other-reports, ability items that can be scored as correct or incorrect, and coding of interviews. The second contribution proposes a whole new approach to the assessment of emotional abilities by using multimedia applications that greatly enhance the ecological validity of the assessment procedures. In the third contribution a plea is made to embed the assessment of emotional intelligence in current emotion theory. Based on a componential emotion perspective theoretically-grounded and empirically-based scoring keys are proposed. Finally, very critical evidence is presented on the context-free assessment of emotional expression, which often forms a key part of emotional intelligence testing. As the four contributions differ with respect to their evaluation of the EI construct, the symposium will end with a discussion on how these recent developments overcome or just confirm earlier criticisms.

Paper 1

"Alternative methods assessing the emotional intelligence of chinese respondents"

Wong, Chi-Sum (The Chinese University of Hong Kong, Hong Kong SAR, China)*
Peng, Kelly (Hong Kong Shue Yan University, Hong Kong SAR, China)
Huang, Emily (Hong Kong Baptist University, Hong Kong SAR, China)

There are severe criticisms on the construct and the measures of emotional intelligence (EI) in late 1990s (Davies, Stankov, & Roberts, 1998). In response to these criticisms, we have been developing four different methods to assess Chinese EI in the past 12 years according to the ability model of EI, i.e., EI is defined as the ability to deal with emotions rather than personality. The first method is the development of a self-report measure. The development process and evidence of reliability and criterion-related validity is reported in Wong and Law (2002). The second method uses other-rated items. That is, raters are evaluating people whom they are familiar with. Evidence of reliability and criterion-related validity for this method is reported in Law, Wong and Song (2004). The third method is to develop test items that have correct versus incorrect options. The development process and evidence for reliability and criterion-related validity is reported in Wong, Law and Wong, 2004 and Wong, Wong and Law, 2007. The final method is to assess Chinese EI level by specific questions used in selection interview. We have gathered reliability and validity evidence for some carefully developed situational interview and behavioral interview questions. Details of the developmental process, and the pros and cons of the above four methods are discussed.

Paper 2

"Multimedia assessment of emotional abilities: Research and development"

Roberts, Richard (Educational Testing Service, USA)*
Schulze, Ralf (University of Wuppertal, Wuppertal, Germany)
Minsky, Jennifer (University of Wuppertal, Wuppertal, Germany)
MacCann, Carolyn (University of Sydney, NSW, Australia)

Research examining emotional abilities (EA) is in the ascendancy, with several target articles in influential journals such as the American Psychologist and the Annual Review of Psychology. However, limitations in extant assessments and the need for alternative measurement approaches are apparent. We discuss the development of two multimedia tests: (1) A situational judgment test (where participants rate a scenario for emotional relevance and salience) and (2) a principal-agent paradigm (where event-emotion contingencies in others have to be perceived and memorized and emotion-behavior contingencies inferred from observed behavior to predict future behavior). In two studies (N=857) these EA assessments were administered to community college and university students across the USA. Study 1 evaluates the psychometric properties, including tests for measurement invariance and examination of subgroup differences (e.g., ethnic groups). We also present the new tests' relationships with emotions measures (e.g., the Mayer-Salovey-Caruso Emotional Intelligence Test), outcome measures (such as GPA and coping with stress), personality (as assessed by the Big Five), and five broad cognitive ability factors (i.e., fluid, crystallized, fluency, spatial, and quantitative ability). Study 2 examines test-retest reliability, along with relationships that the measures share with positive affect, as assessed by the Day Reconstruction Method. Overall, findings suggest that multimedia assessments of EA are reliable and share meaningful relations with (a) crystallized intelligence, (b) emotions measures, and (c) valued outcome variables (e.g., coping with stress). We conclude with a discussion of some limitations and future research that aims to address identified problem areas and extend the multimedia approach.

Paper 3

"Constructing scoring keys for the assessment of emotion knowledge"

Fontaine, Johnny (Ghent University, Belgium)*
Scherer, Klaus (University of Geneva, Switzerland)

A major issue for the viability of the emotional intelligence construct is the scoring key of the ability items. How can one determine the correctness of an answer to a question about emotions? The three main approaches, namely consensus scoring, expert scoring, and target scoring, have been severely criticized. They would not be accepted as valid ways to identify correct answers of classical intelligence items. In the present paper an alternative approach is proposed for one aspect of the EI construct, namely for emotional knowledge. The approach is based on the GRID instrument, which consists of 142 features that operationalize six emotion components (appraisals, expression, subjective experience, bodily reactions, action tendencies, regulation) and 24 emotion terms (for the assessment of meaning) or daily emotional episodes (for the assessment of experiences). It probes the salience of each emotion feature for each emotion word or for each daily episode. In a first study in Belgium, Switzerland, and the UK (Fontaine, Scherer, Roesch, & Ellsworth, 2007), a principal component analysis revealed a robust overall four-factorial structure (pleasantness, potency, arousal, and unpredictability) for the meaning of emotion words. This structure was confirmed in a recent, large cross-cultural study in 30 linguistic/cultural groups from five continents and in a large-scale emotion episode study in Belgium. It will be discussed how the results from the GRID studies can be used to derive theoretically-grounded and empirically-based scoring keys for emotional knowledge items.

Paper 4

"Emotion judgments are relative: Implications for assessing emotional intelligence"

Yik, Michelle (The Hong Kong University of Science and Technology, Hong Kong SAR, China)*
Zeng, Kevin (The Hong Kong University of Science and Technology, Hong Kong SAR, China)

Judging others' emotions is central to daily social interactions and is the basis upon which teachers, parents, lovers, and friends behave. The ability to make accurate emotion judgments was used as a benchmark for distinguishing people diagnosed with schizophrenia or ADHD from controls (Kerr & Neale, 1993; Rapport, Friedman, Tzelepis, & Voorhis, 2002) and for assessing individual differences in emotional intelligence (Roberts, Zeidner & Matthews, 2001). Pertinent to this everyday wisdom are the assumptions that we automatically express our emotions via verbal or nonverbal behaviors and that observers, with minimal efforts, are capable of efficiently decoding the behaviors and correctly judging the emotions expressed by others. In the present study, we examined the mechanism underlying the process of judging others' emotions from the anchoring and adjustment perspective. We showed that judgments of emotions communicated in emotion scripts were influenced by the context for judgment (viz. an anchor). Our results challenge the practice of using emotion judgments as a yardstick to measure emotional intelligence.

Discussant

Grégoire Jacques (Université Catholique de Louvain, Belgium)*

Invited symposium 7 15:15-16:45 LT4

Recent developments of CBT in Japan

Chair

Shigemasu, Kazuo (Teikyo University, Japan)

Symposium Abstract

Computer Based Testing has been finally getting popular in Japan. In this symposium, we will introduce some unique efforts both in terms of theory and practice to promote CBT in Japan.

Paper 1

"Implementing multidimensional item response models for routine computer based testing"

Muraki, Eiji (Tohoku University, Japan)*

Compensatory and non-compensatory multidimensional item response theory (MIRT) models have been constructed and their parameters are estimated by the marginal maximum likelihood (MML) method. Any cognitive tests seem to be hardly unidimensional because their cognitive tasks require various combinations of complex mental functions. However, it is quite difficult to use routinely MIRT models to the standardized testing situations because the MIRT models are built to aim essentially at capturing the interactions between test items and subjects' complex cognitive performances and those interactions can be thought to qualitatively differ among subgroups of subjects, such as their gender and ages. The routine implementation of the MML method is also causing problems because the estimation method needs multiple integrations and their complexity increases exponentially as the number of dimensions is added. In this presentation, the MCMC method is derived to estimate the parameter values of the MIRT models and suggest a reasonable procedure which can be implemented routinely for standardized computer-based testing applications.

Paper 2

"Introduction to the Common Achievement Test System for entering clinical clerkship in Japanese medical schools"

Mayekawa, Shinichi (Tokyo Institute of Techonology, Japan)*

Common Achievement Test was introduced in 2005 in order to assess the students' mastery of the core curriculum before entering clinical clerkship in Japan. The test was developed using IRT models and administered through the network of computers.

Paper 3

"Challenges in developing and operating CBTs in Japan"

Nogami, Yasuko (The Japan Institute for Educational Measurement, Inc., Japan)*
Kobayashi, Natsuko (The Japan Institute for Educational Measurement, Inc., Japan)
Hayashi, Norio (The Japan Institute for Educational Measurement, Inc., Japan)

It has been about eight years since the Japan Institute for Educational Measurement (JIEM) released the CASEC, a computerized adaptive testing system to measure proficiency of English as foreign language. Through Internet, examinees can take the test at any time and any place, and they receive feedback immediately upon completion of the test. The number of examinees has been steadily increasing. In 2009, the test was taken more than 110,000 times. The examinees range widely in background—from junior-high-school students to university graduates and adults in the workforce. The results of the test are used for different purposes and in different contexts; placement in schools, monitoring educational achievements, and so on. The JIEM also has another type of computerized test called CASEC-G. Examinees are required to translate Japanese sentences into English ones, and their writing skills are evaluated. After taking the test, examinees receive some advice on how to improve their performance and they can brush up their writing skills with a tutorial system called CASEC-GTS. In addition, the JIEM will release a new computerized test to assess reading skills of English in April 2010. We would like to introduce these computer based tests developed and operated by the JIEM. We will illustrate some difficulties we encountered in the process, and our efforts to solve them.

Paper 4

"A nationwide listening comprehension test using personal IC players"

Otsu, Tatsuo (The National Center for University Entrance Examinations, Japan)*
Uchida, Teruhisa (The National Center for University Entrance Examinations, Japan)
Ito, Kei (The National Center for University Entrance Examinations, Japan)

A Japanese scholastic standard nationwide examination, called the National Center Test (NCT) is conducted by NCUEE in January every year. All national and local public universities as well as part of private universities make use of NCT. Usually, each university administers its own tests in February of March. Currently, the use of NCT by private universities is supplementary. Usually they assign a small portion of the admissions for NCT applicants, and the rest are assigned to applicants for their own examination. There were more than a half million applicants participated in NCT in 2009. NCT is designed to assess the basic scholastic achievements which applicants have attained in upper secondary high school. NCT 2009 provides tests in 28 subjects in six areas, Japanese language (including Japanese and Chinese classics), geography and history, civics, mathematics, sciences, and foreign languages. Every applicant is not required to take all the six subject areas, but each university designated the subject areas or subjects at its discretion. English test of NCT contains listening comprehension test items in addition to usual paper and pencil test items. The NUCEE conducted the listening test in 2006 at the first time. The listening comprehension test of NCT consisted of 25 short questions, and took 30 minutes. We will introduce our operation of the test, contents of the test items, and its influences on the university admissions in Japan.

Discussant

Zhang, Houcan (Beijing Normal University, China)*

Wednesday, 21st July 2010

Invited symposium 8 08:30-10:00 LT5

The Wechsler intelligence scales cross the globe: Measurement variance and invariance.

Chair

Weiss, Lawrence G. (Pearson, USA)

Symposium Abstract

This symposium is about the measurement invariance and variance of the Wechsler Intelligence Scales. Up to date, most research about the measurement invariance and variance of the Wechsler intelligence scales was conducted using data from the west globe. With the publication of WISC4 and WAIS4 in Taiwan, Hong Kong, and Mainland China, we will report our recent findings about the measurement variance and invariance of the Wechsler Scales using the Asian data. Four papers related to measurement invariance, language effect, Flynn effect, and validity evidence of the Wechsler scales will be reported. The theoretical, practical, and clinical meanings of the results will be discussed within the context of previous studies using the data from the west globe.

Paper 1

"The measurement invariance of WISC4 Hong Kong, Macau, Taiwan, and Mainland China."

Chen, Hsin-Yi (Pearson, USA)*
Weiss, Lawrence G. (Pearson, USA)
Li, Yuqiu (Beijing Normal University at Zhuhai, China)

The purpose of this study was to test measurement invariance of the WISC-IV factorial structure between four regions: Hong Kong, Macau Taiwan, and Mainland China. The structure reported in the US WISC-IV manual (Wechsler, 2003) was used as the hypothesized baseline model. Then, multi-sample analyses were conducted with constraints embedded in a stepwise manner. We tested for invariance on four levels of nested models. Each level had more constraints than the previous one (Meredith, 1993). The first and weakest level was configural invariance. It assumed the overall factor pattern was the same between regions. The second level was testing for weak factorial invariance, also called metric invariance. This model required the magnitude of the factor loadings to be the same between regions. The third stage tested unique variance invariance, which examines whether the test measures the same construct with similar accuracy. Finally, in the most restricted model, the factor covariances were all constrained to be equal across genders. All factor models were tested using covariance matrices. Maximum likelihood was the estimation method chosen because of its robustness and sensitivity to incorrectly specified models. During each step of the analyses, the chi square difference (Δx2) was tested between nested models and suggestions regarding partial measurement invariance were carefully considered and followed. If inadequate fit was detected, fit in the model was improved by including additional parameters identified by the modification index (MI) provided by LISREL. Re-parameterization was examined carefully for meaningfulness.

Paper 2

"Language effects on the performance of WISC4 subtests: Evidence from the U.S., Hong Kong, Macau, Taiwan, and China."

Li, Yuqiu (Beijing Normal University at Zhuhai, China)*
Zhu, Jianjun (Pearson, USA)
Chen, Hsin-Yi (Pearson, USA)

The current study evaluates the language effects on the performance of WISC4 subtests. First, we hope to replicate the results reported by Chen and Zhu (2004) using the WISC4 data from multiple samples. Second, we want to evaluate the language effects on children's performance of the Digit Span subtest. Because the numbers (digits) used in Chinese language are short and simple, they are much easier to memorize and pronounce. As a result, children from Hong Kong, Macau, Taiwan, and Mainland China should show higher digit span forward and backward scores than U.S. peers. Samples matched on parent education, age, and sex were drawn from U.S., Hong Kong, Macau, Taiwan, and Mainland China normative samples. Next, children's performance on the WISC4 Coding, Symbol Search, Digit Span, Matrix Reasoning, Block Design, and Picture Concept subtests were compared across the five matched samples.Preliminary results confirmed the previous finding by Chen and Zhu (2004). Children from Hong Kong, Macau, Taiwan, and Mainland China did significantly better on Coding and Symbol Search subtests than American peers. In addition, children from Hong Kong, Macau, Taiwan, and Mainland China also did significantly better on Digit Span Forward and Backward subtests. On average, U.S. children scored 4.5-4.9 points lower on Digit Span Forward and 0.5-2.2 points lower on Digit Span Backward. The theoretical, practical, and clinical implications of these results on cognitive research, test development, and clinical practice will be discussed.

Paper 3

"Cross-culture comparison of Flynn effect on Wechsler Intelligence Scales"

Zhang, Houcan (Beijing Normal University, China)*
Zhu, Jianjun (Pearson, USA)
Chen, Haipin (Beijing Normal University, China)
Chan, Yat (Educational Psychology Service, China)

For more than two decades, research has shown consistent support for the Flynn effect on Wechsler intelligence scales. However, due to a shortage of data, there are very few studies specifically evaluating the Flynn effect on the Wechsler intelligence scales using Asian samples. The current study will evaluate the Flynn effect on Wechsler intelligence scales using data from Taiwan, Hong Kong, and Mainland China. Data from the four validity studies will be used to evaluate the Flynn effect. Composite scores will be used in the data analysis. If possible, data from the four validity studies will be pooled to increase the statistical power of the current study. The current study will be focusing on the following research questions: (1) Are the Flynn effects observed from these four studies consistent with the expectation set forth by Flynn (1984, 1987), i.e., about a 0.3 increase in FSIQ points per year? (2) Are the Flynn effects observed in the current study invariant from those reported in U.S. edition of the Wechsler manuals? (3) Are the Flynn effects observed in the current study invariant across the four samples? Are there any age trend? (4) Are the Flynn effects observed in the current study invariant across all ability levels? The theoretical, practical, and clinical meanings of the results will be discussed within the context of previous studies that evaluate the Flynn effect on Wechsler scales.

Paper 4

"Evidence of reliability and validity of WAIS4 China."

Zou, Yizhuang (Beijing Huilongguan Hospital, China)*
Wang, Jian (Beijing Huilongguan Hospital, China)

The Chinese adaptation of the U.S. WAIS4 is currently in progress, and the standardization of the instrument will be finished by early 2010. As part of the standardization, the following Chinese samples will be collected: a normative sample, a test-retest sample, an inter-scorers reliability sample, a paper-penciled and digital administration equating sample, and a couple of clinical samples. The current presentation will focus on the flowing psychometric properties of the Chinese WAIS4: (1) Representativeness of the normative sample; (2) Evidence of reliability, such as internal consistency reliability, test-retest stability, and inter-scorer agreement; (3) Evidence of validity, such as exploratory and confirmatory factor analyses, inter-subtest correlations, correlation between the previous and current edition, and the equivalency of paper-pencil and digital administration; and (4) Initial evidence of clinical validity based on a sample of individuals diagnosed with schizophrenia and a sample of individuals diagnosed with mental retardation. The consistency between the psychometric properties of the Chinese WAIS4 and the U.S. edition will also be discussed.

Discussant

"The past, current, and the future of the research on the Wechsler intelligence scales"

Grégoire, Jacques (Université Catholique de Louvain, Belgium)*

The discussion will be focus on the following: (1) A brief review about the previous cross cultural research on the Wechsler intelligence scales; (2) The theoretical, practical, and clinical implications of the four papers presented at the symposium; (3) the direction of the future cross-culture research on the Wechsler scales.

Invited symposium 9 08:30-10:00 LT6

A new generation of DIF studies

Chair

Elosua, Paula (University of the Basque Country, Spain)
Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

Numerous DIF studies have been published in specialized and applied psychometric journals during the last two decades. In addition to the development of statistical procedures for detecting differential item functioning that are highly efficient in spotting problematic items, the research on DIF to date has also focused on applications of DIF analyses in a range of testing contexts. All of this work is critical because of the extent to which DIF analyses are a fundamental part of item analysis. However, it is important to note that any analysis of differential item performance should not be narrowly focused on the detection of DIF: once DIF is detected, the task turns to understanding it, the study of effects of item type on examinee performance, or the study of the practical consequences. It is this idea of extending DIF studies with new methods and approaches that forms the basis of this proposed symposium: A new generation of DIF studies. The new perspective involves multilevel latent models, mixed models, consequences and new robust procedures for the detection of DIF. The symposium consists of four presentations given by researchers from four countries. The first study illustrates a new approach to detecting DIF based on using robust statistics ; the second one uses a simulation to evaluate the effects of factorial partial invariance on group comparisons ; the third and fourth presentations incorporate mixture models to evaluate the presence of latent classes and novel applications of multilevel IRT .

Paper 1

"Robust anchoring and posterior anchoring as procedures for DIF and measurement equivalence"

de Boeck, Paul A. L. (University of Amsterdam, Netherlands)*

An important issue in the process of identifying DIF and also in the process of obtaining measurement equivalence is the choice of anchor items. The basis for this choice is commonly either prior knowledge or iterative purification based on the data. Two alternatives are presented here: (1) robust anchoring, using tools from robust statistics, and (2) posterior anchoring, based on posterior DIF probabilities of the items. The robust approach can be implemented in a parametric way, for example with a robust version of the Raju distance, or in a nonparametric way, for example with marginal proportions correct. The posterior approach requires a mixture model for the items, with a DIF class and a non-DIF class of items. These two alternatives do not require prior knowledge and neither do they make use of iterative purification. They both rely on a one-step statistical procedure. Simulation studies show that their performance is excellent. Apart from their practical use in dealing with DIF and obtaining measurement equivalence, they are also novel IRT approaches in a more fundamental statistical sense.

Paper 2

"The effect of Partial Factorial Invariance on group comparisons"

Elosua, Paula (University of the Basque Country, Spain)*
Zumbo, Bruno D. (University of British Columbia, Canada)

Factorial invariance studies examine the equivalence among factorial structures across groups. Conclusions about partial factorial invariance mean that some of the model parameters (loadings, thresholds, error variances) are different for groups. It is difficult, however, for a researcher to quantify the effects (i.e., impact) of this lack of invariance on subsequent statistical decisions based on group mean comparisons or coefficient alpha comparisons across groups.

Paper 3

"Latent variable mixture modeling as a method to examine sample heterogeneity, and the related problem of DIF"

Zumbo, Bruno D. (University of British Columbia, Canada)*
Sawatzky, Richard G. (Trinity Western University, Canada)
Ratner, Pamela A. (University of British Columbia, Canada)
Kopec, Jacek A. (University of British Columbia, Canada)

We will present an overview of a program of research that applies latent variable mixture modeling (LVMM) to examine the extent to which a sample is homogeneous with respect to a specified statistical model for ordered categorical item responses. Along the way we will evaluate the implications of sample heterogeneity with respect to the latent variable scores, and identify potential sources of sample heterogeneity. As has been shown in the literature, LVMM can be used in conjunction with IRT (i.e., an IRT mixture model) to examine sample heterogeneity, and the related problem of DIF, when relevant group differences are not assumed a priori (Cohen & Bolt, 2005; De Ayala et al., 2002; Mislevy, Levy, Kroopnick, & Rutstein, 2008; Rost, 1990; Samuelsen, 2008; Vermunt, 2001). Our aims are: (a) to share the lessons we have learned about LVMM, its implementation and limitations, and (b) demonstrate how looking at the typically DIF situation from this vantage point allows us to investigate whether there are other variables than the usual manifest variable in DIF studies (such as gender, age, or nationality), or interactions among variables, that distinguish homogeneous groups. Our focus will be typical psychosocial measures such as emotional wellbeing and physical functioning, and the data complexities they present.

Paper 4

"Applications of multilevel IRT models to investigate item type effects"

Zenisky, April L. (University of Massachusetts, USA)
Elosua, Paula (University of the Basque Country, Spain)
Zumbo, Bruno D. (University of British Columbia, Canada)*

This presentation focus on a new multilevel IRT model and on its application to study item type effects which can affect the performance across groups. A multilevel IRT model developed for group-level diagnosis was applied to study data from high school end-of-course examinations. Variability in item difficulty across ethnic groups was investigated in relation to item features associated with content and cognitive process categories. Random effects were attached to each feature type at the group level, and their variability studied across groups. The estimated feature effects were shown to provide a basis for examining cross-ethnic differences for individual features as well as cross-feature differences within individual ethnic groups, as this may be useful for diagnostic purposes. The model was fitted using Markov Chain Monte Carlo procedure by R software.

Invited symposium 10 10:00-11:30 LT5

Development of psychological test in Mainland China

Chair

Zhang, Jianxin (Institute of Psychology, Chinese Academy of Sciences, China)

Symposium Abstract

Four speakers will present separately their papers on the application of psychological measurements such as MMPI and CMHI, and on use of the new techniques such as IAT in developing tests, and on Ethical Code of psychological tests endorsed in Chinese mainland.

Paper 1

"Theory and method of psychological and educational measurement being widely applied in Chinese mainland"

Zhang, Minqiang (Southern China Normal University, China)*

In Chinese mainland, the candidates of any test are numerous because of a huge population. The number of university entrance exam takers has reached 10 million, and the number of candidates participating in the entrance exams for postgraduate schools has increased to 1.2-1.5 million. Moreover, there are a large number of candidates in other examinations, such as the judicial examination, the qualified doctor practitioner examination, the accountant qualification test, the civil service examination, and even in tests for foreign candidates, such as HSK. The increasing application of test and the improvement of demand for test organization has strongly pushed the theoretical and application research of psychological and educational measurement in Chinese Mainland. A psychometric committee has been established under Chinese Psychological Society; under the Chinese Society of Education, there is a branch for educational measurement and statistics, with approximately 1,000 members. These professionals active in all kinds of field are prompting the theoretical and application studies in psychometrics, and great success has been achieved in theory and application of CTT, GT and IRT. Plenty of scales with conformance to psychometric rules have been widely used for different kinds of population, which play an important part in improving the mental health of Chinese people as well as preventing the mental disease.

Paper 2

"The 2008 revision of the Chinese Code of Ethical Use of Psychological Tests"

Gan, Yiqun (Beijing University, China)*
Che, Hongsheng (Beijing Normal University, China)

In response to the rapid increase of application and abuse of psychological tests in China, the Psychometrics Division of Chinese Psychology Society (CPS) made major revisions to the Chinese Code of Ethical use of Psychological Tests in 2008. Comparing to the earlier version in 1992, the rules were reorganized to define more specifically the responsibilities of test users and the rights of test takers. New items concerning the test users' qualification and the validation of instruments were added. The test users are recommended to use only those psychological tests approved or registered by the CPS. In addition, a number of points relevant to the respect for test takers' rights and privacy were stated more explicitly. The current code provides the psychological test users in China clear guidelines for ethical decision making in their work.

Paper 3

"The arena of mental health measurements in Mainland China: From SCL-90 to CMHI"

Chen, Zhiyan (Institute of Psychology, Chinese Academy of Sciences, China)
Wu, Zhenyun (Institute of Psychology, Chinese Academy of Sciences, China)
Huang, Zheng (Institute of Psychology, Chinese Academy of Sciences, China)*
Guo, Fei (Institute of Psychology, Chinese Academy of Sciences, China)

The translation and normalization of many foreign mental health measurements has been done since 1980's. In the past 30 years, SCL-90 has been the most used mental health measurement in college and hospital settings. Other often used instruments in college settings and in hospital settings differed. The former were UPI, EPQ, and 16PF, the later were SAS, SDS, HAD, HAMD, BDI, etc. As the most used mental health measurement in mainland China, SCL-90 has been criticized for its improper application in community sample, inability to identify "negative symptoms", and so on. To provide an instrument more applicable in non-patient sample, Chinese Mental Health Inventory (CMHI) was developed to measure individual's level of mental health with psychological concepts rather than psychiatric symptoms. CMHI has five dimensions, including emotion experience, self-evaluation, interpersonal capacity, cognitive efficacy and adaptiveness. Aside from the attempt to assist diagnosis of several mental disorders, follow-up mental health service system after measurement was also provided for CMHI.

Paper 4

"The clinical application of the MMPI in Mainland China"

Wang, Li (Institute of Psychology, Chinese Academy of Sciences, China)*
Zhang, Jianxin (Institute of Psychology, Chinese Academy of Sciences, China)

The Minnesota Multiphasic Personality Inventory (MMPI) was first introduced into mainland China in the early 1980s. As an objective personality test with sound psychometrical properties, the MMPI rapidly became one of the most popular assessment instruments in clinical setting in mainland China. It was widely used as a screening or an aided diagnosis tool in variety of populations, and has been demonstrated to have useful clinical applications. However, given that the short history of the MMPI in mainland China, its clinical application is mainly limited to basic clinical scales and a few content scales. Therefore, more studies should be conducted to further validate new developed content scales, additional scales, and special scales for promoting their clinical application in Chinese population.

Discussant

Zhang, Houcan (Beijing Normal University, China)*

Comments on the above four speakers' presentations in particular, and on Chinese psychological tests in general will be provided.

Invited symposium 11 10:00-11:30 LT6

Assessment models for monitoring learning

Chair

Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

The symposium will discuss on the models for monitoring teaching and learning in three countries: Denmark, Hong Kong and New Zealand. The three systems will be reviewed and discussed in terms of their influences on learning and teacher autonomy, the stakes associated with assessments, the types of assessments used, the levels of aggregation of data from these assessments, and how data are used.

Paper 1

"National tests in Denmark – CAT as a pedagogic tool"

Wandall, Jakob (Danish Ministry of Education, Denmark)*

Testing and test results can be used in different ways. They can be used for regulation and control, but they can also be a pedagogic tool for assessment of student proficiency in order to target teaching, improve learning and facilitate local pedagogical leadership. To serve these purposes tests have to be low stake. In Denmark, to ensure this, test results are made strictly confidential by law. The only test results that are made public are the overall national results. Because of the test design (Rasch-model), results are directly comparable, which gives an enormous potential for monitoring added value and developing new ways of using test results in a pedagogical context. The presentation gives the background and status for the development of the Danish national tests, describes what is special about these tests (IT-based, 3 tests in 1, adaptive, etc.), how the national test are carried out and what is tested. Futhermore, it is described who are allowed to know the results, what kind of response is given to the pupil, the parents, the teacher, the headmaster and the municipality and how the results can be used by the teacher and headmaster.

Paper 2

"Alternatives to external standardized assessments: Hong Kong example"

Hamp-Lyons, Liz (University of Hong Kong, Hong Kong SAR, China)*

In this presentation I will 1) describe the school-based assessment system that has been introduced across Hong Kong secondary education to assess the English speaking skills of all students; 2) describe how this classroom assessment data is used to report student level data for educational planning and region-wide accountability; 3) discuss how and to what extent this school-based assessment supports learning in the classroom and contributes to teacher professional development.

Paper 3

"Assessment models for monitoring learning: New Zealand"

Hattie, John (The University of Auckland, New Zealand)*

New Zealand has a recent history of self-managed schools with many freedoms to make decisions about teaching and assessment. There are many options for them to choose. The session outlines the options available in an on-line assessment package (asTTle) which includes Teacher customised, comprehensive, computer adaptive, interview, and attitude assessment. Feedback is immediate to teachers and students in the form of visual reports, and while they can be used for many purposes the major use is to monitor teaching and learning.

Paper 4

"Comparing and contrasting models for monitoring learning in three countries"

Ercikan, Kadriye (University of British Columbia, Canada)*

This presentation will review, compare and discuss models for monitoring teaching and learning in three countries that will be presented in the first part of the symposium: Denmark, Hong Kong and New Zealand. The three systems will be reviewed and discussed in terms of their influences on learning and teacher autonomy, the stakes associated with assessments, the types of assessments used, the levels of aggregation of data from these assessments, and how data are used.

Discussant

von Davier, Alina A. (Educational Testing Service, USA)

附錄二、第七屆國際測驗年會照片

7th CONFERENCE OF THE INTERNATIONAL TEST COMMISSION
PHILIPS
Theme: Challenges and Opportunities in Testing and Assessment in a Globalized Economy
July 19-21, 2010
(July 18, 2010 Pre-conference workshops)
Hong Kong
www.itc2010hk.com

2010 年 ITC 會議

會議報到

Scientific Program
State-of-the-Art Address 2
Keynote Address 5
Invited Symposium 11
Special Session 5
Diamond Sponsor Session 2
Symposium 25
Oral paper session 20 (109 Papers)
Poster session 7 (135 Posters)

年會開幕式

Co-organized by
International Test Commission
Department of Psychology of
The Chinese University of Hong Kong
7th CONFERENCE OF
THE INTERNATIONAL
TEST COMMISSION
WELCOME
TO
HONG KONG

開幕合影

新任主席

前任主席與香港中文大學校長合影

The 7th Conference of
the International Test Commission
ITC
July 19 – 21, 2010
The Chinese University of Hong Kong

新任主席

Evaluating test quality as users and writing manuals as authors: Two sides of a coin 工作坊

Establishing the ITC Guidelines on quality control in scoring, analysis and reporting of test scores 工作坊

會議研討(一)

Reasons for the Second Edition
Reasons for the Second Edition

會議研討(二)

會議研討(三)

International Test Commission
Guidelines for Test Adaptation,
Second Edition (Draft)
Ronald K. Hambleton
Center for Educational Assessment
Univ. of Massachusetts Amherst, USA

會議研討(四)

Applications of Multilevel IRT Models to Investigate Item Type Effects

會議研討(五)

ISO
International Organization for Standardization
Questions:
1. What are the key potential benefits of an ISO approach to psychological assessment and what are the possible downsides?

會議研討(六)

18
12
11
The impact of instant feedback during
unsupervised online

海報展覽(一)

21
22

海報展覽(二)

贊助商展覽(一)

贊助商展覽(一)

茶敘

餐會

與國內學者合影

閉幕式