行政院國家科學委員會專題研究計畫 成果報告

結構化教師甄試口試之相關系列研究（III）
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2413-H-004-006-
執 行 期 間 ：95 年 08 月 01 日至 96 年 12 月 31 日
執 行 單 位 ：國立政治大學教育學系

計 畫 主 持 人：胡悅倫
共 同 主 持 人：余民寧
計畫參與人員：博士班研究生-兼任助理：陳世芬、李仁豪

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 96 年 12 月 12 日

行政院國家科學委員會補助專題研究計畫 ∨ 成 果 報 告 □期中進度報告

結構化教師甄試口試之相關系列研究（Ⅲ）

計畫類別：∨ 個別型計畫 □整合型計劃
計畫編號：NSC 95-2413-H-004-006-
執行期間：95 年 8 月 1 日至 96 年 12 月 31 日

計畫主持人：胡悅倫
共同主持人：余民寧
計畫參與人員：兼任助理陳世芬(博士班學生)、兼任助理李仁豪(博士班學生)

成果報告類型(依經費核定清單規定繳交)：∨精簡報告 □完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
∨出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：

中 華 民 國 九 十 六 年 十 二 月 十 二 日

壹、前言

教師甄試口試之目的在選擇有潛力且具專業知能的優秀教師，因此，在口試過程中，口試題目即扮演著關鍵性的角色。但是，面對這些關鍵性的題目，社會大眾對它的瞭解微乎其微，而且相較於筆試題目，極少有人願意針對口試題目做深入的探討。觀之目前心理計量研究的重點，大都放在探討筆試題目的信度、效度、抑或鑑別度等相關議題，甚至有人認為口試題目所在乎的不是題目本身的內容，而是在口試過程中，應試者的反應或是人格特質。然而，口試題目是否為一值得探討的議題呢？相信讀者很有興趣瞭解此一問題的答案。

口試在人事甄選中一直扮演著舉足輕重，幾乎無可取代，此乃因其能測量到其他方法不可得到的特質。這些特質大致上可區分成三大類別（Gatewood & Field, 2001）：（1）人際關係；（2）組織公民行為；（3）專業工作知能。在人際關係方面，如社交能力及語言流暢度。無庸置疑地，良好的人際關係及語言能力，對個人在人際網路的建立及組織目標的達成，具有關鍵性的影響力，也間接及直接影響個人的生涯發展。其次，在組織公民行為方面，是指個人在組織中，願意多付出心力，以維護社群良好的互動或發展。此種特質，也說明一個人的獨立性、誠實、穩定性及堅持度。此種特質通常可在口試中,與應試者討論其工作習性、工作完成度、或工作環境後瞭解得知。除了人際關係能力及組織公民行為之外，口試最為特別的地方，在於它可以成功的測量出專業工作知能。關於工作上的專業知識，若其答案是簡短而制式的，那麼筆試將是較佳的選擇。但若涉及複雜的專業行為，如工廠中的機械操作，或是在學校中教學方法的運用，則口試將會是較好的選擇。研究者以為，假如口試委員少，或者應試者少，題目也少，那麼口試題目被研究的價值性則小。畢竟口試在乎的是個人特質、人際關係、組織公民行為、專業工作行為等特質。但若現在某種口試有上千題，或者數千題的口試題目，則其代表的意義就全然不同了。

目前研究者因從事教師甄試的相關議題研究，在同一年度已蒐集到4901題的口試題目。其中86.59%（4244題）乃透過台北縣聯合教師甄試委員會中78位口試委員，詢問將近兩千多名的應試者，當中每位應試者的口試時間僅有6-10分鐘，而78位委員花了將近八小時的口試時間所得到的資料。換言之，同一天內有78個人對同一個主題——甄選優良教師，進行腦力激盪所得到的結果，研究者認為這是一個非常珍貴的資料。對於一個心理測驗研究者而言，這無疑是個寶物。套句徐志摩的話「數大便是美」。相信許多讀者皆同意這個觀點。而這些題目的總合，是否能反應出人事心理學對口試的期待，亦即這些口試題目真能測到「人際關係」、「組織公民行為」及「專業工作知能」嗎？這些疑惑皆讓口試成為一個有趣且具價值的研究議題。

另外，研究者認為這些口試題目到底透露出何種訊息？口試委員心中都在想些什麼？是一樣或是不一樣的議題？或是個人的，抑或是環境的？是教室中的班級經營，抑或是當前熱門的教育政策問題？是行政服務抑或是教學能力？是教育專業知識抑或是個人生涯規劃？而哪些類別是最常被詢問的的問題？而哪些類別是與教師工作最有關聯的呢?而哪些類別是應試者最難回答的呢?相信許多讀者亦跟研究者一樣，對這些問題背後的答案有所期待。

目前，在口試議題的研究上，研究者已從第一階段4901題中凝聚組合成十六個次類別，其中包含一類為無法歸類，但在分類的過程中，卻碰到以下幾個瓶頸：(1)每一類下的題目

數仍然太多太雜，難以瞭解。例如，「教學能力」有649題，「自我介紹」有675題，「班級經營」有853題等，其中有完全重複的，有部分雷同的，亦有完全不同的。(2)目前這些口試題目均是以文字的方式出現，究竟該如何以質量並重的研究方法，勾勒出一個完整的圖像，可供敘說、討論與研究? (3)到底這些口試題目背後隱藏著何種理念？而這些理念又如何聚斂成一個可用以描繪未來優秀教師的完整圖像？

為解決上述這些瓶頸，且為國內口試研究奠定完整的理論基礎，研究者決定應用概念構圖的方法----結構化概念形成法（structured conceptualization），針對口試題庫進行系統性研究，故本研究之具體目的如下：

一、應用概念構圖法，具體呈現口試題目的圖像。

二、透過口試題目圖像的描述，深入瞭解口試委員腦海中的抽象知識表徵，以幫助日後口試甄選活動的進行。

貳、文獻探討

一、口試

口試乃是教師甄試的重要關鍵，其結果往往決定一位教師的錄取與否。而一位教師的晉用卻又關係著千萬學子的學習生涯。欲求孩子能在學校快樂成長、享受學習，甚至卓然有成，師資的良窳乃是一件不容忽視的課題。是故，口試的重要性則是不言而喻。在口試的過程，所有的應試者無不卯足全力地，希望在短短的口試中脫穎而出，雀屏中選。而在決勝的關鍵時刻中，應試者必須在口試委員面前將其背景合宜地展現出來，對於口試委員所提之問題，應試者須極力展現其專業的程度。換言之，在口試過程，應試者必須將其知識與能力轉化到工作上，同時能展現個人在人格與專業上的優勢，期能達到工作上的要求(Murray, 1999; Ream, 2000)。

教師甄試面談應試者必須能展現解決問題的能力，統合專業的技能及永續智能成長的可能性。就某程度而言，經由面談的過程應試者越能回答越多的問題，則口試委員則可做出越有效的決定（Sdetsky ＆ Pell,1980）。面談並不只是在蒐集應試者的知識，更重要的是了解其運用知識的能力（Roberts, 1987）。有學者認為，一位應試者必要讓人感受到充滿活力、值得信賴、願意與他人合作，甚至是願意勤奮工作的（Braun, 1990）。

通常大部分的候選人都是進入試場的前幾分鐘就已經被決定，而其後的時間大概只是再度的確認當初的決定（Braun, 1990; Murray, 1999; Sharp ＆ Sharp, 1997）。其實對大部分的校長而言，通常在口試之前對某些職務的人選都有個特定的圖像（Sharp ＆ Sharp, 1997）。因此除了對候選人的再度確認之外，口試委員在整個口試的過程就是在尋找該工作適合的候選人特質。

在口試的技巧上，80/20 的黃金原則顯然非常適用。Kirkwood 與 Ralston (1999)強烈的建議。應試者要能在口試期佔去大部分的時間，才是致勝的關鍵。也就是說，在整個口試期間應試者最好能佔用 80％的談話；而口試委員則佔用 20％的時間。換句話說，對於口試委員的問題，應試者越能回答或愈能提供詳細的訊息，顯然被錄取的機會就比較大。

筆者以為，口試最重要還是要回歸到口試是否有效的議題上，這也是社會大眾最關心的焦點。而在效度上的研究，發現口試結構化程度有助於提升口試的效度。「結構化面試」

英文譯為”structured interview”，亦稱為標準化（standardized）、指導性（guided）、系統性(systematic)或是組型式(patterned)面試。結構化面試之所以能提升面試效度，在於面試委員對於決定該問何種問題與如何評價應試者的反映有一套完整架構。由後設分析研究更證實非結構面試的效度大約是在 0.14 到 0.30 之間，但是，一旦面試結構化後，其效度可增加至 0.35 至 0.62（Campion, Palmer, & Campion, 1997； Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; Marchese & Muchinsky, 1993; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989），亦即結構式面試的效度是非結構式面試的二倍。

而在結構化的口試中，人事心理學在面談甄試的研究成果豐富；觀之國外對於結構化面試的文獻（如 Burnett, Fan, Motowidlo, & Degroot, 1998；Campion,Palmer, &Campion, 1997；Huffcutt &Woehr, 1999；Janz,1982；McDaniel, Whetzel, Schmidt, &Maurer, 1994; Motowidlo, Carter, Dunnette, Tippins, Werner, Burnett, & Vaughan, 1992；Wiesner & Cronshaw, 1988; Wright, Lichtenfels, & Pursell, 1989），其中 Campion, Palmer & Campion（1997）提出的十五項影響結構性面試的因素整理的最為詳盡，而其中與面試題目有關的因素如後：

1.面試題目擬定

結構性面試的題目必須奠基於重要事件(critical incidents)的「工作分析」。所謂「工作分析」即該工作領域的專家學者使用問卷調查或專家座談方式，分析特定工作的重要細節，使能對工作的內容有一提綱挈領的瞭解。透過「工作分析」的步驟，面試委員便能掌握特定工作必須具備的能力，亦能針對這些必要的能力或重點發展出面試問題形式及內容，藉由這些問題的詢問，使應試者表現出這些問題相對應的能力或特質，以利面試委員做出反映應試者真實性的評分，達到選才之目的。「工作分析」不僅允許面試委員獲得應試者行為有關的工作樣本的資訊（Dipboye & Gaugler, 1993），也強化了與工作有關之訊息數量，使重要訊息均能被納入考量。

2.應試者均被詢問相同之問題

面試結構化程度取決於面試委員的提問標準化。而提問標準化即為面試委員針對每位應試者依相同的順序詢問相同的問題，若面試委員使用相同的語言詢問，更能提高結構化的程度。雖然高度結構化的面試效度很理想，但實際運作卻很困難，因為難保應試者不會洩露相關訊息給後續應試者，然而，使用相同的問題是將面試從「討論」轉向「科學性測量」的重要關鍵（Campion, Palmer, & Campion, 1997）。因此，如何克服實際困難，使用相同問題是提升面試效度的必要步驟（Schriesheim, Solomon, & Kopelman, 1989）。

3.即興問題或後續追問的處理

一般而言，面試著重應試者與面試委員的動態互動歷程，在面試過程中，面試委員有時會不按既定的問題發問，或者會針對應試者先前的回答，額外或臨時提問一些與工作內容無關的問題，目的可能是為了獲得更多有用的訊息，以利決策評分。但是這些互動歷程卻容易讓面試委員在訊息收集上產生偏誤（Dipboye, 1994; Jelf, 1999）。為了避免不必要的訊息進入面試中，並提升面試的結構性，對於即興問題應該避免，而後續追問也應該扣緊主題，同時每一位應試者都能獲得相同的追問與澄清的機會。

4.問題形式與內容

早期研究（Janz, 1982; McDaniel et al., 1994; Motowidlo et al., 1992; Motowidlo, Dunnette, & Carter, 1990）將面試問題形式歸納為六種：如「情境式問題」、「過去行為問題」、「背景式問題」、「與工作知識有關的問題」、「真實工作的模擬」「心理特質問題」等。研究者亦發現：以「情境式問題」及「過去行為問題」的面試效度最佳，而「心理特質問題」的面試效度最低（Janz, 1982； McDaniel et al., 1994）。其中情境問題的效度建立在目標或意圖與未來行為的關係上（Locke & Latham, 1984）；而過去行為問題的效度通常是假定過去行為是未來行為良好的預測因子（Mumford & Stokes, 1992）；心理特質問題十分模稜兩可，卻足以讓應試者表現出令人喜愛的態度或避免表現出缺點（Campion, Palmer, & Campion, 1997）。Taylor 與 Small (2002)後設分析進一步顯示，過去行為問題的平均效度高於情境問題（0.63 vs. 0.47）。

5.面試的時間與題數

一如測驗長度對測驗信度的影響，當測驗長度愈長時，信度愈好；而信度愈佳，效度也會隨之提高。同理，面談的時間與題數的增加時，面談的效度也會增加。面談長度是一個基本，但又常被忽視的因素，在合理的時間下，「面談時間」與「面談題數」的增加導致結構性較佳，因為較長的面談可獲得有關應試者較多的資訊（Campion, Palmer, & Campion, 1997）。過去的研究並未特別重視面談長度對面談結構性的影響，Campion、Palmer 與 Campion（1997）所收集到的近 200 篇有關面談研究的文獻中，只有 38 個研究報告面談的時間，其範圍在 3 至 120 分鐘間，而平均面談時間是 38.75 分鐘；並且只有 14 個研究報告面談的問題題數，其範圍在 4 至 34 題間，而平均題數是 16.50 題。但沒有一篇研究對所選擇的面談長度做解釋。至於，到底要多少時間及多少題目，目前尚無定論，有待學者進一步研究。

是故，由面試題目的擬定、詢問方式、後續問題的處理，形式與內容，甚至是面試的時間與題數，都與口試結構化程度息息相關。而要提升口試結構化程度，方能增進口試的效度。然而要改進口試結構化程度的要項中，最困擾口試委員的工作，莫過於面試題目的擬定（胡悅倫，2007）。目前台灣教師甄試口試，為了保密的原則，面試委員通常在口試的前幾天才接到通知。到時間集合，頂多會前討論，並沒有所謂的面談訓練，或工作分析以擬定口試問題（胡悅倫，2007）。

理想上，口試問題的擬定，最好是由工作分析得之。但，工作分析乃為一浩大的工程。其實也有人注意到，有關口試的題目在職場上的專家，就經常寫許多的面試問題，如：經理或是面試委員（Janz, 1982; Latham et al., 1980; Orpen, 1985; Roth &Campion, 1992）。但大部分的論文並沒有很詳細的說明口試題目是如何擬定的，所以顯然口試題目大部分都是由研究者自己撰寫（Palmer & Campion, 1997）。而筆者以為，這正是學術研究上的缺口。目前，在論述教師甄試的研究中（Mondak, 2004; Hinderman, 2004），研究者在從事口試問題研發過程，的確是經由文獻探討，先瞭解所謂優秀教師，或稱效能教師（effective teacher）的可能性定義，而後依此定義再由研究者自行研發口試題目。是故，此種方式，一方面無法脫離他人思考的窠臼；另一方面，口試題目缺乏真實性，無法瞭解專家或口試委員對工作信念的看法，而在教師甄試，即是專家對教師此一職務的看法。是故，本研究分析教師甄試口試題目的內容，或許也可以解決目前國內教師甄試中所遇到的困境。

目前國內教師甄試文獻在通俗性期刊或教育雜誌方面，其重點大多聚焦在公平性的問

題上。這些文章非嚴謹的研究報告或實證性論文，但多以口試委員或應試者觀點發表，多少能反應口試委員及應試者的心聲（如：石弘毅，2000；宋慶瑋，2003；陳坤德，1999；陳維貞，2001；楊素菱，2004；蘇鈺琦，2004）。對教師甄選評分的標準、公平性及口試委員的聘任方式提出質疑（石弘毅，2000；宋慶瑋，2003；陳維貞，2001；蘇鈺琦，2004），並認為口試和試教的評分流於主觀，建議有效增加評量者的客觀性（楊素菱，2004）；亦有認為甄選過程粗糙，甄選結果不具公信力的（蘇鈺琦，2004）。這些通俗性期刊或教育雜誌的文章，因非嚴謹的研究報告或實證性論文，故其客觀性及說服力仍嫌不足。

國內針對教師甄選之相關研究頗多，研究的範圍廣及教師甄選的制度、甄選程序、甄選標準、甄選成效、甄選主辦單位、甄選委員遴聘、應試者資格...（李居憲，1998；呂祥義，2002；吳福春，2002；李慶宗，2001；李燕綺，1999；胡文仲，2004；姜智棟，2003；張喬媚，1998；葉連祺，1997；劉秀蓮，2004；劉佳鵑，2003；蔡秉修，2002；蘇婉芬，2004；蘇鈺琦，2004）。只是這些研究多就教師甄選的制度面進行剖析及檢討，鮮少針對於甄選的過程中最受質疑的口試進行探究；更遑論對口試題目做分析，了解面試委員的背後知識表徵，及其教育意涵。是故，本研究之重要性不言而喻。

二、結構化概念形成法

一般而言，在計畫或評鑑某件專案的過程中，最困難的問題或步驟，可能是面對如何將原本是南轅北轍的各種原始構想予以明確地概念化，以作為後續活動的遵循依據。換言之，在計畫的過程中，研究者會期望能夠將構成計畫的主要目的和目標、需求、資源、和量能或其他元素等，予以形成概念並表徵出來；在評鑑的過程中，可能需要將有關聯的方案或處理方法、樣品、情境、測量、和結果等，也能形成概念並具象表徵出來。在此，所謂的概念形成（conceptualization），即是指將各式各樣的想法、創意觀點、預感及其表徵方式，以某種明確、客觀的形式清楚表達出來的意思（Trochim, 1985; Trochim & Linton, 1986）。

本研究在此所採用的結構化概念形成法，即是一種運用概念構圖（concept mapping）方法將眾多原始構想予以結構化形成的最佳工具（Trochim, 1989a; 1989b; 1989c）。這種結構化的概念形成過程，不僅可以提供一般性的概念架構（conceptual framework），以導引團體成員在專案計畫與評鑑初期即可形成理論與概念，更可以作為一種研究方法或研究工具，以幫助發展或進行適當的問題解決。概念構圖在這方面的應用，是結合概念圖及其他足以表徵與解釋各式各樣觀點的圖形表徵方式（程序如同 Novak & Gowin（1984）所述的概念構圖方式），並且增加卡片歸類（card sorting）與評分程序（Rosenberg & Kim, 1975），以及兩種多變量統計分析技術（multivariate statistical methods）──即多元度量法（multidimensional scaling）（Davison, 1983; Kruskal & Wish, 1978）和群集分析（cluster analysis）（Anderberg, 1973; Everitt, 1980），共同應用在分析這種圖形資料所提供的訊息和結果上，以具體的操作型定義步驟方式和衍生出來的概念表徵過程，來達成結構化的概念形成目的（余民寧, 1997; Trochim & Linton, 1986）。

目前，國內運用結構化概念形成法於創新研究專案的例子，已屢見不鮮，例如：吳政達、郭昭佑（1997）應用到國小教科書評鑑標準的建構上、郭昭佑（2000）應用到國中校務評鑑指標的建構上等。由這些應用成果可知，在在顯示結構化概念形成法具有十足的應用潛力。本研究即是基於此理念，擬採用作為本研究的核心方法。

參、研究方法

一、研究對象

由於教師甄試辦理時程各縣市的重疊性高，故在蒐集資料上無法全面施測，故本研究以立意抽樣方式，選擇台北縣與高雄縣聯合教師甄試應試者做為主要的受試者，再輔以一所台北市自辦完全中學教師甄試應試者人員。其中台北縣的聯合教師甄試考場分設於五個學校，分別為永平高中、錦和高中、永和國中、中和國中、福和國中，而台北市自辦教師甄試則以萬芳高中為主。是故，本研究樣本涵蓋自辦、公辦兩種方式，以及北高兩個地區。希望藉此方式，得以對不同辦理方式、不同地區的教師甄試口試題目有整體性的瞭解。

本研究資料蒐集以問卷調查方式為主，請應試者在口試結束後，將口試的題目記錄下來，共收到 4901 道題目，形成教師甄試口試題庫。而其中台北縣聯合甄選題目共佔 86.59％（4244 題）。之後，由研究者、專家教授，以及三位具有十年教學經驗的在職國/高中教師，以教育現場實務經驗，先選定一所學校的口試題目，歷經十次的共同討論研究，形成共識分析，建構十六項主類別。而第十七類為無法歸類，其中無法歸類者為語意不清或是答非所問者，合計為 177 題，佔總題數的 3.61% 。之後，依循相同分類之標準，再進行其他試場口試題目之歸類。各類別的名稱及題數詳如表 1：

表 1　教師甄試口試題庫十六大類組次數分配

項次	主類別	口試題目數	百分比
1.	班級經營	853	17.40%
2.	自我介紹	675	13.77%
3.	教學能力	649	13.24%
4.	行政服務	463	9.45%
5.	過去表現	390	7.96%
6.	教育理念	367	7.49%
7.	教育政策	317	6.48%
8.	輔導知能	309	6.30%
9.	專業知識	239	4.88%
10.	實習教師	177	3.61%
11.	學校環境	165	3.37%
12.	個人價值	42	0.86%
13.	人際關係	31	0.63%
14.	生涯規劃	30	0.61%
15.	資源支援	11	0.22%
16.	休閒生活	6	0.12%

表 1　教師甄試口試題庫十六大類組次數分配

項次	主類別	口試題目數	百分比
1.	班級經營	853	17.40%
2.	自我介紹	675	13.77%
3.	教學能力	649	13.24%
4.	行政服務	463	9.45%
5.	過去表現	390	7.96%
6.	教育理念	367	7.49%
7.	教育政策	317	6.48%
8.	輔導知能	309	6.30%
17.	無法歸類	177	3.61%
	總計	4901	100%

自表 1 可知，最常發問的口試題目類別之次數，前三項分別是：「班級經營」853 題（17.40％）、「自我介紹」675 題（13.77％）、「教學能力」649 題（13.24％）。其它中間類別依序為：「行政服務」463 題（9.45％）、「過去表現」390 題（7.96％）、「教育理念」367 題（7.49％）、「教育政策」317 題（6.47％）、「輔導知能」306 題（6.24％）、「專業知識」239 題（4.88％）、「實習教師」177 題（3.61％）、「無法歸類」177 題（3.61％）、「學校環境」136 題（2.77％）。而使用次數不到 1％的類別則依序為：「個人價值」42 題（0.86％）、「人際關係」31 題（0.63％）、「生涯規劃」30 題（0.61％）、「校園安全」29 題（0.59％）、「資源支援」11 題（0.22％）、「資源支援」11 題（0.22％）、「時事問題」3 題（0.06％）。

再者，從口試題目的分類折線圖可看出三點結論（如圖 1）：

1.同一區域但不同試場（口試委員）的題目分配彼此很相似，從台北縣五個公辦教師聯合甄試口試題目可以得知。

2.公辦與自辦的口試題目分配，大致而言，非常相似(除班級經營的類別外)，可由台北縣五個公辦試場與台北市萬芳高中自辦試場得知。

3.不同區域的口試題目分配非常相似，可從台北縣、臺北市及高雄縣得知。

由上述可知，口試問題的分佈情形是相當相似的。也就是說，口試問題的類型不受到試場（口試委員）、甄選方式、區域等因素的影響，所以大家最關注的議題大都集中在班級經營、自我介紹、教學能力、行政能力、過去表現等。最不常被重視的議題是「校園安全」、「資源支援」、「休閒生活」、「時事問題」等。此一重要的研究結果確立口試題庫的可行性。未來若欲增加題目，以充實題庫的豐富性，建議以現場錄音口試題目的方式，較能完整記錄口試題目的原貌。不過口試事關考生權益，口試現場採集方式，知易行難。是故，能對整個考場做實際問卷調查已實屬不易。

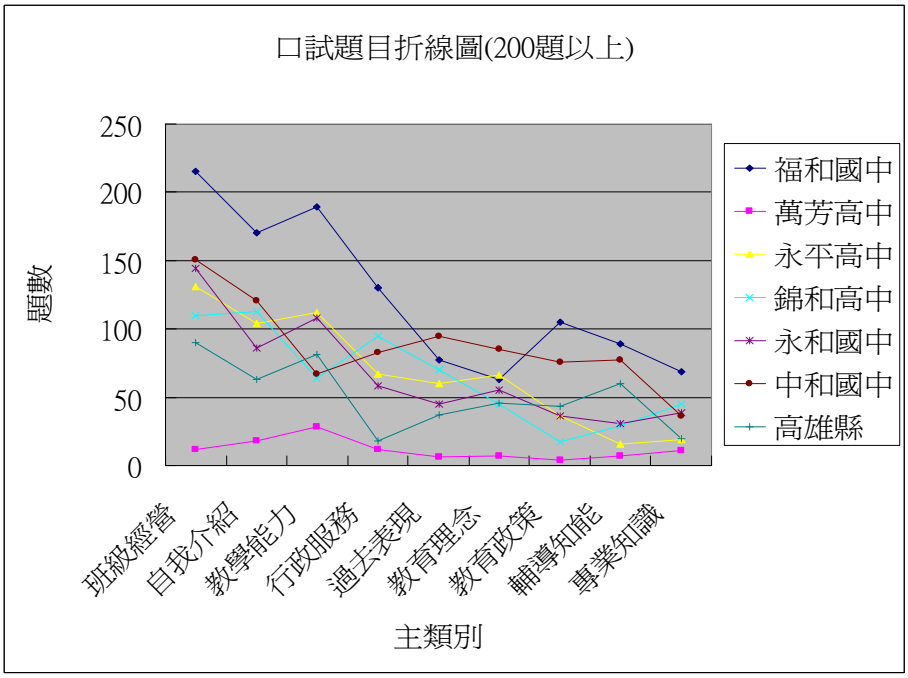

圖 1 口試題目的分類折線圖

二、結構化概念構圖之分析

基於上述之研究目的與文獻分析，本研究擬以結構化概念形成法進行兩階段的分析，茲扼要陳述如下（余民寧，1997）：

階段一：以概念構圖法，瞭解十六大主要類別，所聚斂出的抽象知識表徵，以描繪未來優秀教師的具體圖像。

階段二：以概念構圖法，分析各類別中典型代表之口試題目，並了解各類別所隱含的教育理念。

(一)準備

在這個階段中，主要的工作項目有二：

1.挑選參與者

本研究選取 2 位大學教授、3 位國中校長、2 位國中主任(或組長)、7 位不同科別之國中資深教師、5 位不同科別之國中初任教師，以及 1 位國中家長等共 20 名利害關係人（stakeholders），作為本研究之團隊成員。在教授、國中校長、主任及資深教師方面，均具有至少十五次以上的口試評審經驗，初任教師則具有參加口試的經驗，而家長則為參與觀察者。在此，選擇此團隊成員的標準，是在促使團體成員性質達到最大異質化為目的，並確保成員的多樣性及代表性。為使此參與者能對本研究慎重其事，每位參與分類口試題目的人員，本研究稱為審題委員。

2.篩選典型代表題目

扣除無法歸類或語意不清之題目，就目前的有效口試題目可粗略分為十六大類組。擬請不同的專家，進行兩大步驟，篩選出每一大類組的題目，控制在一百題以內，以便後續作業的處理。

步驟一：刪除重複題目。即視每種題目單獨出現的情形，將類似的題目組成一群，本研究大約粗略分成十六類組。此一步驟，已由研究者及三位資深教師完成。

步驟二：就每一組群中，撰寫一題概念化題目做為代表。就一群組的題目，選出最能代表某一概念化的題目，作為典型代表題。此一步驟請五位資深（均具 20 場以上口試經驗）專家處理之。在資料檢核上，將每一概念化題目，以小組討論的方式審查，逐一確認概念化題目的合理性。以下將舉例說明如何將原始題目歸納成一組概念的題目，如:原始題為「如何將六大議題融入教學？」、「如何將兩性議題融入教學？」、「如何將環境議題融入教學？」、「如何將七二水災融入教學？」、「如何將自己的課外經驗融入課程？」，而其概念化題目為「如何將○○議題(六大議題、社會議題、個人專長等)融入教學？」

(二)陳述典型代表題

為了考量將來資料分析時的電腦容量及人為力量一次所能處理的數量，我們可以限定以一百個不同典型代表題為最大的處理量。接下來，分別將這些典型代表題直接登錄在 3 x 5 吋的卡片上，每張卡片僅能登錄一個典型代表題，以方便後續的資料處理。由於口試題目的來源，是以應試者考完口試後，在試場外填寫的。故，某些句子不通順的情況，擬請國文老師一名，將所有的句子，在不更改原意下，潤飾成為明確易懂的語句。

（三）典型代表題的結構化

在這個階段中，主要的工作項目有二：

1.典型代表題的歸類與資料登錄

一旦完成上述步驟，接下來，即可參照 Rosenberg & Kim（1975）所建議的非結構性卡片分類程序（unstructured card sorting procedure）來進行分類整理工作。

首先，研究者要求每位審題委員「以自己感到最有意義的方式」，將上述登錄在卡片中的每個典型代表題，進行歸類整理。在歸類整理過程中，審題委員必須遵守下列幾項規則的限制：

A.每個典型代表題只可以被歸到某一類之中（亦即，每個典型代表題不可以同時被歸到二類裡）；

B.所有的典型代表題不可以全部被歸在同一類之中；

C.也不可以將所有的典型代表題各自獨立歸成一類（雖然如此，但其中有某些典型代表題可以單獨歸成一類）。

除了遵照上述的限制外，審題委員愛怎麼歸類就怎麼歸類，但每種歸類行為都必須依

照審題委員自己覺得最意義的方式而進行者。因此，審題委員可以發現歸類的方式有許多種，但每一種歸類方式對他們而言，都是有意義的。也許，審題委員需要多練習歸類幾次，才能決定最後他所滿意的歸類結果。

一旦每位審題委員完成他的歸類工作，所完成的結果必須加總起來。這有兩項作法如下：第一，將每位審題委員的歸類結果，登錄到與所歸類的典型代表題數目相等的方格或矩陣中，該矩陣中的細格內元素值只有兩種：0 與 1；其中，1 表示某位審題委員將某個行與列的典型代表題歸在同一類別上的意思，而 0 即表示某位審題委員不將某個行與列的典型代表題歸在同一類別的意思。第二，資料登錄之後，係以「二元化對稱近似矩陣」方式輸出，每一位審題委員有一張輸出的矩陣資料表，然後，將個別的歸類矩陣資料加總起來，以獲得一個整體的近似矩陣資料表。

這個最後完成的近似矩陣資料表，即為某個概念領域的相關結構，因為它可以提供所有審題委員如何歸類所有關點的訊息。在該矩陣中的元素數值愈大，即表示愈多的人以相同的方式將兩個不同典型代表題歸為同一類，它隱含著這些典型代表題在某種程度上是概念相似的；反之，矩陣中的元素數值愈小，即表示較少的人以相同的方式將兩個不同典型代表題歸為同一類，它即隱含著這些典型代表題在概念上是較不相同的。

2.典型代表題的評定

其次，根據「回答上的困難度」、「與工作表現的關聯度」兩個問項，要求受試者針對每個典型代表題進行四點評定量表式的評定，以彰顯每個典型代表題的重要性、優先性。對每個典型代表題而言，至少我們可以獲得評定結果的團體平均數，或其他可進行描述統計分析的訊息資料。另外，每一個題目的代表題數，亦可作為一種評定依據。只是此一評定，可由原始資料計算得知，無需審題委員再費時評定之。故，本研究中各典型代表題的相對重要性，係由下列三項指標的測量決定：「題目與優秀教師工作表現之關聯性」、「代表題數」、「回答的困難度」。

（四）典型代表題的表徵

這個步驟的任務主要在進行概念圖的計算，有三項工作要做：

1.進行多元度量法分析

針對上述步驟所得的近似矩陣資料表，進行二向度的非參數多元度量法（nonparametric multidimensional scaling）分析。這項分析結果將可獲得一個估計點圖（point map）。

2.進行群集分析

接著，根據上述多元度量法分析後，以所求得每個估計點在二向度構圖上的座標值，作為進行群集分析的輸入資料，並採用華德氏（Ward's）的階層群集分析法（hierarchical cluster analysis），找出少數的幾個群集（clusters），形成一些群集圖（cluster map），再看看每個群集圖內到底包含哪些典型代表題，以方便後續的解釋工作。

3.計算估計點圖和群集圖的平均評定值

一旦獲得上述分析後的兩種構圖，一為估計點圖，另一為群集圖；接著，便是計算每個典型代表題的平均評定值，稱為「估計點評定圖」(point rating map)，以及每個群集的平均評定值，稱為「群集評定圖」(cluster rating map)。算法很簡單，只要將所有審題委員在每個典型代表題上的四點評定量表評定值加總起來，再除以審題委員的總人數，即可得估計點的平均評定值；若將每一群集內的各個典型代表題的所有評定值加總起來，再除以該群集內的典型代表題數目，即可獲得群集的平均評定值。

（五）概念圖的解釋

在這個階段中，主要的工作即是針對下列項目進行解釋：

（1）典型代表題清單：即由所有審題委員所提出的原始典型代表題的彙整清單。

（2）群集清單：即經由群集分析後，被歸到同一群集內的所有典型代表題清單，研究者將試圖針對每個群集分別予以命名，以達以簡馭繁的解釋效果。

（3）估計點圖：即經由多元度量法分析後所得的估計點構圖，研究者可分別予以標號，以方便解釋。

（4）群集圖：即經由群集分析所歸類而成的群集構圖。

（5）估計點評定圖：即已計算出平均評定值的估計點評定圖，根據這個圖的平均值大小，研究者可以看出到底是哪個典型代表題最受審題委員們的重視。

（6）群集評定圖：即已計算出平均評定值的群集評定圖，根據這個圖的平均值大小，研究者可以看出到底是哪個群集是構成這整體概念形成的核心概念。

根據上述幾項解釋重點，研究者再與審題委員共同討論，以檢驗圖中比較接近的群集是否比較疏遠的群集，在概念上較近似於所有審題委員所認可者。如果不是，研究者可與審題委員共同找出造成這種差異的進一步可能解釋。幾經修正討論後，終可獲致一個經過命名的群集圖，足以代表整個結構化概念形成過程的主要概念架構，同時，也是概念構圖過程的最基本結果。

肆、結論與討論

一、典型代表問題清單及其重要性

教師甄試口試的目的，即是在甄選未來具有潛力的優秀教師。因此，本研究根據結構化概念形成法的分析結果，將可供教師甄試口試使用的原始典型代表題目，經過群集分析的結果，呈現在如表 2 及圖 2 裡。

表 2 所示，即是群集分析的群數凝聚的過程。由表中「係數」一欄內的數值顯示，選取 2 或 4 群都可能是個理想解；不過選擇 2 群的解，比較無法進行有意義的解釋，並且也與實際的教育概念不符，因此，本研究斟酌圖 2 所示的樹狀圖分佈之後，決定選取 4 群的解，以符合理論涵義，並能做有意義的詮釋。因此，本研究決定選擇這十六大類別口試題

目可以形成四大面向區塊，作為本研究群集分析的較佳解。

表 2　群集分析中的群數凝聚過程

階段	組合集群		係數	先出現的階段集群		下一階段
	集群1	集群2		集群1	集群2	
1	1	2	1.000	0	0	3
2	14	16	.988	0	0	4
3	1	3	.947	1	0	7
4	14	15	.930	2	0	12
5	10	12	.913	0	0	9
6	5	7	.911	0	0	13
7	1	9	.888	3	0	10
8	4	11	.883	0	0	10
9	8	10	.857	0	5	13
10	1	4	.797	7	8	15
11	6	13	.763	0	0	12
12	6	14	.671	11	4	14
13	5	8	.492	6	9	14
14	5	6	.422	13	12	15
15	1	5	.264	10	14	0

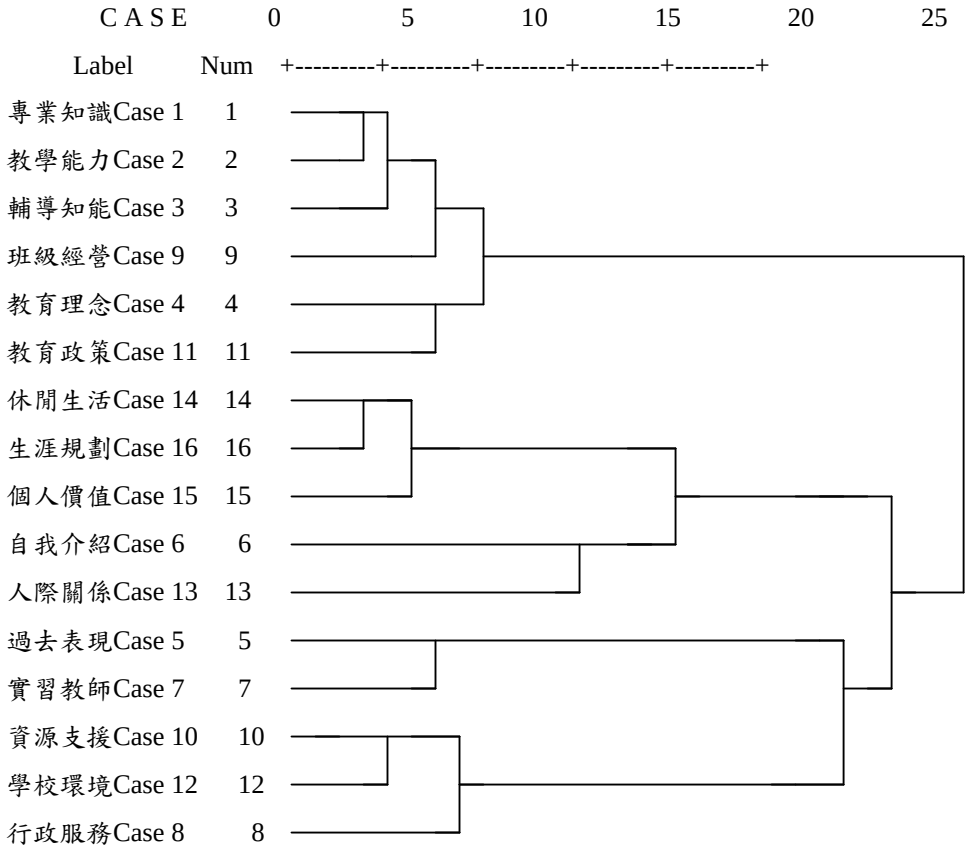

圖2 群集分析的樹狀圖分佈

接著，將這經由群集分析所得的四個群集，拿給 20 名利害關係人去針對各群集加以命名討論之後，教師甄試口試用的問題共計可形成四大面向（dimensions）的典型問題，這四

大面向分別取名為：(1) 教師專業知能面向（含「教學能力」、「班級經營」、「專業知識」、「教育理念」、「輔導知能」、「教育政策」等類別)；(2) 公民組織行為面向（含「行政服務」、「資源支持」、「學校環境」等類別)；(3) 個人先備知能面向（含「過去表現」、「實習教師」等類別)；(4) 個人核心價值面向（含「人際關係」、「個人價值」、「生涯規劃」、「自我介紹」、「休閒生活」等類別)。其中，各群集的命名及其所含類別，如表 3 所示。

表 3　教師甄試口試題目的四大面向關聯性及其回答困難度

口試面向	平均面向關聯性	平均面向難度	口試類別	類別題數	各主類典型代表題	平均類別關聯性(SD)	平均類別難度(SD)
教師專業知能	2.98	2.06	班級經營	853	42	3.16(.79)	2.02(.88)
			教學能力	649	73	2.99(.86)	2.02(.90)
			輔導知能	309	27	2.91(.85)	2.15(.90)
			教育理念	367	34	2.89(.87)	1.99(.93)
			教育政策	317	16	2.78(.91)	2.30(.89)
			專業知識	239	41	2.76(.91)	1.98(.92)
個人先備知能	2.85	1.44	實習教師	177	9	2.95(.88)	1.37(.59)
			過去表現	390	14	2.81(.87)	1.47(.71)
個人核心價值	2.8	1.67	自我介紹	675	12	2.84(.91)	1.67(.85)
			人際關係	31	16	2.67(.97)	1.98(.97)
			個人價值	42	20	2.65(.92)	1.69(.82)
			休閒生活	6	5	2.52(.86)	1.18(.50)
			生涯規劃	30	6	2.39(.97)	1.50(.77)
公民組織行為	2.72	1.95	行政服務	463	23	2.83(.82)	1.97(.93)
			資源支援	11	7	2.73(.82)	2.20(.94)
			學校環境	165	11	2.40(.91)	1.87(.98)

註：1.關聯性量表為四點量表，「1 到 4 分」各代表「非常不相關、不相關、有相關、非常相關」。

2.困難度量表為四點量表，「1 到 4 分」各代表「不困難、有點困難、很困難、極困難」。

3.平均類別關聯性及平均類別難度以典型代表題為計算基礎。

4.平均面向的關聯性及困難性則是以各口試類別題作為基礎之加權平均數。

根據這四大面向的典型問題，審題委員逐一評比每一口試題目在與「優秀教師工作表現之關聯性」、「回答的困難度」等指標上的關係，並在加權平均計算之後，獲得如表 3 所示的結果。表 3 所示，即為四大面向與其內含十六個類別的平均關聯性與平均難度指標值。若從該題口試題目與「優秀教師工作之關聯性」角度來看，由於本研究採用四點評定量表，結果顯示教師甄試口試用的問題所形成的四大面向其平均數均大於量表分數的期望值（即 2.5)，亦即，大部分的教師認為此四個面向與「優秀教師工作表現之關連性」皆有高度相關。因此，若以「優秀教師工作表現之關聯性」為判準，則教師甄試口試時，應該優先考量「教師專業知能面向」，之後考量「個人先備知能面向」、「個人核心價值面向」、「公民組織行為面向」等問題。若從該題口試題目在「回答的困難度」角度來看，結果顯示教師甄試口試用的問題所形成的四大面向其平均數均小於量表分數的期望值（即 2.5)，亦即，大部分的教師認為此四個面向問題的困難度均不至於太難而無法回答。此外，若以「回答的

困難度」為判準，則教師甄試口試時應該以「教師專業知能面向」、「公民組織行為面向」的問題為口試的核心問題，其次才是「個人核心價值面向」與「個人先備知能面向」兩個面向的題目。由此可見，教師甄試在進行口試時，應該包含「教師專業知能面向」、「個人先備知能面向」、「個人核心價值面向」、和「公民組織行為面向」四個面向的問題為口試的核心問題。

在實務應用上，可以參考表 1 所示，進行如下的建議事項。當口試委員欲詢問問題時，在有關「教師專業知能面向」的問題上，可以依序詢問參加甄試者有關其「班級經營」、「教學能力」、「輔導知能」、「教育理念」、「教育政策」、和「專業知識」等類別的題目，因為這些類別的題目與優秀教師的工作表現最有關係；在有關「個人先備知能」的問題上，則可以詢問參加甄試者有關其「實習教師」與「過去表現」等類別的題目，因為這些類別的題目與優秀教師的工作表現也是最有關係的；而在有關「個人核心價值面向」的問題上，則可以詢問參加甄試者有關其「自我介紹」、「人際關係」、「個人價值」、「休閒生活」、和「生涯規劃」等類別的題目，因為這些類別的題目與優秀教師的工作表現也是有關係的；而在有關「公民組織行為面向」的問題上，則可以詢問參加甄試者有關其「行政服務」、「資源支援」、和「學校環境」等類別的題目，因為這些類別的題目與優秀教師的工作表現也是有關係的。其次，在上述十六個類別的口試問題中，每一類別裡至少有 6 題（即「休閒生活」類別）至 853 題（即「班級經營」類別）的題目可供參考詢問，口試委員只要根據每一題目與「優秀教師工作表現之關聯性」的重要程度，依序或隨機抽取問題來詢問均可。

此外，另一個在實務應用上的策略，即是口試委員也可以考慮這些口試題目在「回答的困難度」上的觀點，逐一依序詢問「教師專業知能面向」、「公民組織行為面向」、「個人核心價值面向」、「個人先備知能面向」等面向的問題。以下則依據每個面向裡各類題目的困難度，建議詢問甄試者問題之順序：當在詢問「教師專業知能面向」的問題時，口試委員可以依序詢問參加甄試者有關「教育政策」、「輔導知能」、「班級經營」、「教學能力」、「教育理念」、和「專業知識」等類別的題目（其中「班級經營」與「教學能力」優先順序一致）；在詢問「公民組織行為面向」的問題時，則可以依序詢問參加甄試者有關「資源支援」、「行政服務」、和「學校環境」等類別的題目；在詢問「個人核心價值面向」的問題時，則可以依序詢問參加甄試者有關「人際關係」、「個人價值」、「自我介紹」、「生涯規劃」、和「休閒生活」等類別的題目；而在詢問「個人先備知能」的問題時，則可以依序詢問參加甄試者有關「過去表現」與「實習教師」等類別的題目。其次，在上述十六個類別的口試問題中，每一類別裡至少有 6 題（即「休閒生活」類別）至 853 題（即「班級經營」類別）的題目可供參考詢問，口試委員只要根據每一題目在「回答的困難度」上的程度，依序或隨機抽取問題來詢問均可。

在整個研究過程，本人覺得最興奮的事情是，雖然口試委員沒有受過人事心理學面談理論的訓練，但經過研究以後，發現目前的結果，非常符合人事心理學中對面談的期待。簡言之，經過研究，這些口試題目真的能夠測到筆試所不能涵括的面向，如「人際關係」、「公民組織行為」及「專業工作知能」（Gatewood & Field,2001），並且在教育領域發展出更細緻的具體內容。

讀者可以了解到所謂的「人際關係」，可以相對應於本研究中「個人核心價值」的面向，對於教師工作的範疇中，「個人核心價值」包含「休閒生活」、「生涯規劃」、「個人價值」、「自我介紹」、「人際關係」。因此之所謂的「個人核心價值」，除了人事心理學

提到「人際關係」，又包括一些更廣泛的議題。換言之，教師甄試中，個人核心價值的面像是比單一「人際關係」更完整。而就「公民組織行為」的面向，在教師工作的領域裡，則包含「資源支持」、「學校支持環境」、「行政服務」，相較於人事心理學所談到的面談中「公民組織行為」的特質，讓人更了解到教師領域中所須公民組織行為的特質。而在「專業工作知能」上，本研究所整理的「教師專業知能」面向共包括「專業知識」、「教學能力」、「輔導知識」、「班級經營」、「教育理念」、「教育改革」，也就是說，在教育職場上，教師除了須具有一般的專業知識以外，還需要了解輔導學生的知識，熟悉班級經營的技巧，隨時注意教育政策的改變。

雖然經結果分析「整個」口試題目，在廣度（四大面向）及深度（各面向下包含的內容）均有不錯的發現。但是筆者之見，口試最大的敗筆乃是每場口試的時間只有 10~15 分鐘，而且給每位應試者只有 3~5 題（胡悅倫， 2007），就要決定一位教師的錄用與否，實在是一件危險的事情。相較於平均面談時間 38.75 分鐘（Campion, Palmer, & Campion, 1997），則相去甚遠。在緊迫的口試時間的限制下，口試的內容必定非常侷限，而這對口試的效度必然大打折扣，至少連內容效度都要受到嚴重的考驗。未來最好在口試之前能對口試面向有所共識，使得口試無論在「教育專業知能」、「個人核心價值」、「個人先備知識」甚至「公民組織行為」上都能有合理的分配，使得口試更臻完善。

二、典型代表問題的圖像

概念構圖的主要功能，是以可目視的具體圖像，使讀者瞭解整個教師甄試口試所包含的教育意義，並顯示出其各個面向之間的關係及其相對重要性。如圖 3 所示，我們可以瞭解到，在教師專業中，如「教學能力」、「班級經營」、「專業知識」、「教育理念」、「輔導知能」、和「教育政策」等，是一組概念（即「教師專業知能面向」）；而「行政服務」、「資源支援」、和「學校環境」，則為另外一個概念聚落（即「公民組織行為面向」）；而「過去表現」與「實習教師」，則又是另一概念區塊（即「個人先備知能面向」）；另外，「人際關係」、「個人價值」、「生涯規劃」、「自我介紹」、和「休閒生活」等，則又是另一個概念集群（即「個人核心價值面向」）。其中，各個區塊的高度，即表示與優秀教師工作表現的關聯性，而區塊面積，即代表題數的多寡；因此，區塊高度愈高者，代表其與教師工作表現的關聯性愈高，而面積愈大者，則表示其可被用作口試的題目數量愈多。

另外，由圖 3 所示亦可瞭解到，與優秀教師工作表現關聯性最高的口試題目類別，亦即是工作關聯性平均數最高之前五項，分別是：「班級經營」、「教學能力」、「實習教師」、「輔導知能」、「教育理念」等；而最常被問到的題目類別（亦即是題數最多者），則分別是：「班級經營」、「自我介紹」、「教學能力」、和「行政服務」等；但也可以知道，「自我介紹」雖是最常被問到的口試題目類別，但卻與優秀教師工作表現的關聯性只居中間程度而已，而「實習教師經驗」雖然是較少被問到的口試題目類別，但與優秀教師工作表現的關聯性卻是較高的；而對參加甄試的人而言，「教育政策」、「資源支援」、「班級經營」、「教學能力」、「輔導知能」、「行政服務」等口試題目類別，卻是較難以回答的問題，這亦顯見與「教師專業知能面向」與「公民組織行為面向」有關的題目，都是比較難以回答的問題。

所以說，概念構圖在整個研究問題的具體圖像呈現上，是一種非常淺顯易懂、有效、且簡潔的溝通工具。由概念構圖所隱含的涵義可知，讀者將可以瞭解到：（1）教師甄試口試所應該包含的領域為何；（2）各領域之間的彼此關係及其相對的重要性為何；以及（3）更重要的是，它可以用圖形來表徵所欲研究的概念涵義。

有關教師甄試實証性的研究非常有限。與本研究最相近的研究有兩篇博士論文：1.Mondak（2004）所撰寫的「在教師甄選的影響因素」；2.Hindman（2004）所撰寫的「效能教師與教師甄選的連結---教師甄選的發展草案」。本研究與上述兩篇文章共同之處，均是在討論「在教師甄試中，如何選出優秀或有效能的教師？」；換言之，二者均在討論「教師甄試口試」與「優秀教師」（或稱效能教師）之間的關係，試圖希望經由前者的程序能得到後者的結果。

在 Mondak 則經由文獻探討方式先將所謂效能教師的特質分成四大組群（教學、人格、對學生的態度，及教師資格），擬出 35 道題目，以了解學校行政主管與教師在教師甄選上，對各個效能教師特質看法上的相似程度。

而 Hinderman 亦採同樣的文獻探討的方式。首先，瞭解文獻中「效能教師」的看法，找出五大面向，如人格特質、班級經營、教學組織（organization for instruction）、教學傳送（instructional delivery）及評量。其次，並依此擬 84 道題目。最後，對 300 名實際參與甄選的校長，依「最不滿意」至「典範」等四個類別，評等以上 84 題關鍵性的效能教師特質。誠如 Hinderman 在研究限制所述，即使翻遍效能教師的文獻資料，仍舊沒辦法給效能教師下一定明確的定義。而 Stronge（2002）亦提出同樣的看法：教師的效能是一被籠統模糊提出定義的概念。是的，效能教師或優秀教師乃是一個籠統的概念；但，愈籠統、爭議，則愈是需要用更多元的角度切入；而又因其在教師甄選口試的關鍵性，則愈是需要被關切。是故，本研究從真正面試委員的題目著手，以質量並重的概念構法，聚斂出優秀教師或效能教師的圖像，以期提供一個更完整而豐富的觀點。

筆者以為，相較於過去的研究，本研究有兩項優勢：1.在研究方法上，有別於傳統方式，由上而下，先由別人的觀點，即從文獻探討中，瞭解對效能教師的看法，然後進行調查。如此容易陷在別人對優秀教師（或效能教師）思想窠臼中，無法有創新。事實上，亦證明其研究結果較無個人的洞見。而本研究則是由下而上，先由調查面試委員所問的題目，集合成十大類，然後經由概念構圖聚斂出四大面向。是故，此種資料礦採（data mining）的確讓結果不一樣。2.在研究成果上，本研究結果所聚斂出效能教師或優秀教師，其內容更具教師的生態的觀點，強調教師與環境之間的關係，例如「組織公民行為」的面向。此面向則不見於效能教師的特質相關文獻中。而在效能教師特質中，最常被提及的「教師專業知能」與本研究的面向最為相似。但，除了「輔導知能」、「班級經營」、「教學理念」…等教室中教師教學能力外，本研究也期待一位好的老師，應對環境中的「教育政策」有自己的看法。另外，Mondak 與 Hinderman 雖然亦提到一位效能教師該具有「人格特質」與本研究的「個人核心價值」相似，但本研究的內容更完備。例如，本研究在這個面向包括「個人價值」、「生涯規劃」甚至是「休閒生活」。是故，本研究在研究方法及研究結果的確略勝一籌。

綜合上述，目前的研究資料豐富而多元，故，教師甄試口試題目當然值得研究！但部分的總合是不是等於全部呢？換句話說，是不是問了這些問題後，一個完整的優秀教師圖像就出現了呢？這個問題則有待教育哲學、教育社會學，甚至是教育心理學等，相關教育領域的學者，反思此一深刻的議題。

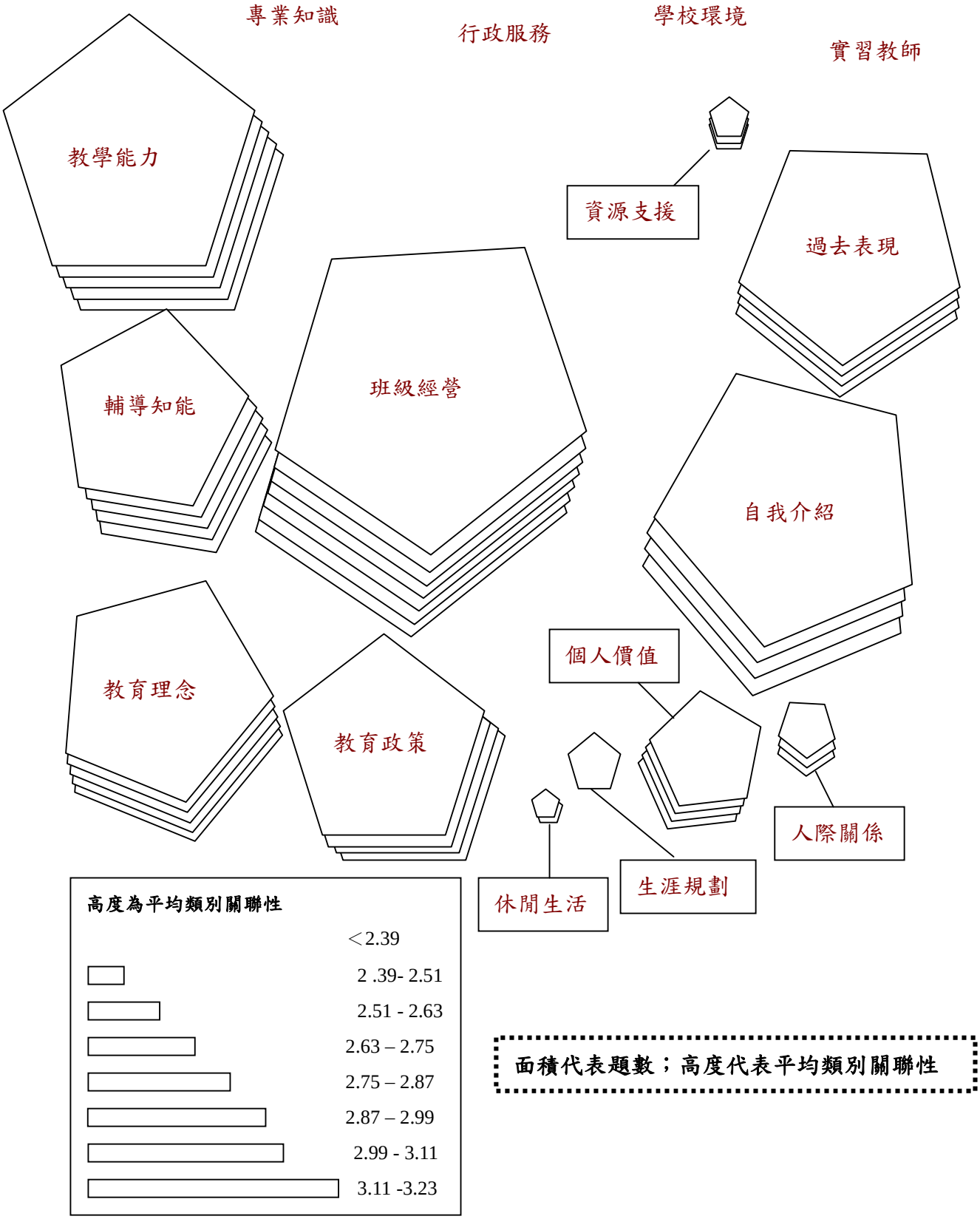

圖 2. 教師甄試口試題目概念構圖之圖示法

參考文獻

石弘毅（2000）。如何辦好教師甄選。**師友，400**，44-47。

余民寧（1997）。**有意義的學習：概念構圖之研究**。臺北：商鼎。

李居憲（1998）。**國民小學實施教師甄選制度之研究**。國立屏東師範學院國民教育研究所碩士論文，未出版。

呂祥義（2002）。**我國小學教師甄選成效之研究**。暨南國際大學教育政策與行政研究所碩士論文，未出版。

吳政達、郭昭佑（1997）。概念構圖法在國民小學教科書評鑑標準建構之應用。**教育與心理研究，20 期(下冊)**，217-242。

吳福春（2002）。**國中教師甄選相關問題研究—以台南縣市為例**。臺南師範學院教師在職進修教育行政碩士學位班碩士論文，未出版。

宋慶瑋（2003）。教師甄選的愛恨情仇。**師友，436**, 32-35。

李慶宗 （2001）。**國民中學教師甄選制度之研究**。國立臺灣師範大學教育研究所碩士論文，未出版。

李燕綺（1999）。**國中教師選聘過程之研究**。國立臺灣師範大學公民訓育研究所碩士論文，未出版。

胡文仲（2004）。**國民小學職前教師對教師甄選制度之意見調查研究---以雲林縣為例**。國立嘉義大學國民教育研究所碩士論文，未出版。

胡悅倫（2007）。結構化教師甄試口試之初步調查。**教育心理學報**，審稿中。

胡悅倫、陳世芬、呂秋萍（2007）。教師甄試面試結構化問卷之編製。**測驗年刊，印刷中**。

姜智棟（2003）。**國民小學教師甄選之研究－以雲林縣為例**。南華大學非營利事業管理研究所碩士論文，未出版。

郭昭佑（2000）。概念構圖法在評鑑指標建構上之應用--以國民中學校務評鑑指標建構為例。**教育政策論壇，3 卷（2 期）**，173-203。

陳坤德（1999）。高職教師甄選經驗談。**職教園地雜誌，88**, 24-25。

陳維貞（2001）。教師甄選經驗談。**國教之友，53(2)**, 15-18。

張喬媚（1998）。**國小教師甄選制度之研究**。臺南師範學院國民教育研究所碩士論文，未出版。

楊素菱（2004）。談教師甄選現況及其改進之道。**師說，183**, 5-8。

葉連祺（1997）。我國中小學教師甄試之研究。**教育資料文摘，39(6)**, 44-64。

劉秀蓮（2004）。**國民小學教師甄選制度實施現況及其改進意見之研究**。屏東師範學院教育行政研究所碩士論文，未出版。

劉佳鵑（2003）。**國民中學教師甄選與遷調制度之政策評估－以嘉義市國民中學為例**。國立中正大學政治學研究所碩士論文，未出版。

蔡秉修（2002）。**國中生活科技教師甄選之研究**。國立臺灣師範大學工業科技教育研究所碩士論文，未出版。

蘇婉芬（2004）。教師甄選作業調查結果摘要及因應策略。**中國統計通訊，15(11)**, 19-26。

蘇鈺琦（2004）。談國小教師甄選制度。**教師之友，43(**5), 81-89。

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Braun, J. (1990). Getting a job: Perceptions of successful applicants for teaching positions. *Action in Teacher Education, 12*(2), 44-54.

Burnett, J. R., Fan, C., Motowidlo, S. J., & Degroot, T. (1998). Interview notes and validity.

Personnel Psychology, 51, 375-396.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655-702.

Davison, M. L. (1983). *Multidimensional scaling*. New York: John Wiley & Sons.

Dipboye RL, Gaugler BB.(1993). Cognitive and behavioral processes in the selection interview. *In Schmitt, N., & Borman, W. C, Associates (Eds.) ,Personal selection in organizations (pp.135-170). San Francisco: Jossey-Bass*

Dipboye RL.(1994). Structured and unstructured selection interviews：Beyond the job-fit model**.** In Ferris GR（Ed.）,*Research in personnel and human resources management：Vol.12*（pp.79-123). Greenwich, CT：JAI Press.

Everitt, B. (1980). *Cluster analysis* (2nd ed.). New York: John Wiley & Sons.

Gatewood, R. D., & Field, H. S. (2001). *Human Resource Selection*. Orlando, FL: Harcourt College Publication.

Hunter, J. E., & Hunter, R. F. (1984). The validity and utility of alternative predictors of job performance**.** *Psychological Bulletin, 96,* 72-98.

Huffcutt, A., & Arthur, W. Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184-190.

Huffcutt, A. I., & Woehr, D. J. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior, 20,* 549-560.

Hindman, J. L. (2004). The connection between qualities of effective teachers and selection interviews: The development of a teacher selection interview protocol. *Dissertation Abstracts International* (UMI No. 3118184).

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67,* 577-580.

Jelf, G. S. (1999). A narrative review of post-1989 employment interview research**.** *Journal of Business and Psychology, 14*(1), 25-58.

Jennifer Lilliston Hinderman（2004）. *The connection between qualities of effective teachers and selection interviews：The development of a teacher selection interview protocol*. The college of William and Mary in Virginia.

Kirkwood, W. G., ＆ Ralston, S. M. (1999). Inviting meaningful applicant performances in the employment interviews. *The Journal of Business Communication, 36*(1), 55-76.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. J*ournal of Applied Psychology, 65*, 422-427.

Locke, E. A., & Latham, G. P. (1984). *Goal-setting: A motivational technique that works.* Englewood Cliffs, NJ: Prentice-Hall.

Mondak. (2004). Influences on Teacher Selection (Virginia Polytechnic Institute and State University, 2004). *Dissertation Abstracts International, 65* , 201.

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment, 1,* 18-26.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied*

Psychology, 79, 599-616.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., & Vaughan, M. J. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77,* 571-587.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Murray, J. P. (1999). Interviewing to hire competent community college faculty. *Community College Review*. *27*(1),41-56.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette ;& L. M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol.3* (2nd ed., pp. 61-138). Palo Alto, CA: Consulting Psychologists Press.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge, London: Cambridge University Press.

Orpen, C. (1985). Patterned behavior description interviews versus unstructured interviews：A comparative validity study. *Journal of Applied Psychology,70,*774-776.

Ream, R., (2000). Why are manhole covers round？*Information Today, 17*(5),26-27.

Roberts, J., (1987). Standardizing the process: How to make the most of teacher interviews. *NASSP Bulletin, 71*, 103-108.

Roth, P. L, & Campion, J. E. (1992). An analysis of the predictive power of the panel interview and pre-employment tests. *Journal of Occupational and Organizational Psychological Bulletin, 65* ,51-60.

Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research, 10*, 489-502.

Sadetsky , I., ＆Pell, A. R., (1980). *Interviewing and selecting elementary school teachers and administrative personnel.* Huntington, NY: Personnel Publications.

Sharp, H., & Sharp, W. (1997). *From field experience to full-time teaching：Letting teachers know how to face interviews and what to expect on the job.* Paper presented at the Annual Meeting of the Association of Teacher Educators, Washington, D. C.

Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement, 13,* 19-32.

Stronge, J. H. (2002). *Qualities of effective teachers.* Alexandria, VA: Association for supervision and Curriculum Development.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75,* 277-294.

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9,* 575-604.

Trochim, W. M. K. (1989a). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12,* 1-16.

Trochim, W. M. K. (1989b). Concept mapping: Soft science or hard art? *Evaluation and Program Planning, 12*, 87-110.

Trochim, W. M. K. (1989c). Outcome pattern matching and program theory. *Evaluation and Program Planning, 12,* 355-366.

Trochim, W. M., & Linton, R. (1986). Conceptualization for evaluation and planning. *Evaluation and Program Planning, 9,* 289-308.

Wiesner, W. H., & Cronshaw, S. F. (1988）. The moderating impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61,* 275-290.

Wright, P. M., Lichtenfels, P. A., & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology, 62,* 191-199.

計劃成果自評

一、教師甄試口試用的題目，具有四大面向，十六大類別的題目可資使用。這四大面向，代表著一位優秀教師所應該具備的四大能力領域；每一個能力領域內，各具有兩至六個能力類別；每一個能力類別內，各具有題數多寡不一的口試題目。這樣的口試題目，可以自成一套標準化的口試題庫，足供各級學校爾後在辦理教師甄選活動時，作為口試問題的參考之用。

二、對於口試題庫的建立，若欲讓口試題目能對不同試場(口試委員)、甄選方式、區域等因素之分配考驗，未來研究應提出更具體之量化數據，方能更有說服性。

三、致於口試題目在效度之建立，除採用橫斷式研究法對不同地區及不同試場之題項分析其相似度與異質性外，未來研究可朝縱貫式研究法，比較優秀及劣質教師後續追蹤，方能判定前置口試題目之鑑別效度。

四、由於這四大能力領域、十六種能力類別的口試用題目，均已獲取其與優秀教師工作之關聯性的指標值，因此，它們可以被分類及排序，並製作成優秀教師評鑑檢核表（evaluation checklist for the outstanding teachers），以供學校作為年度評鑑教師行為表現的工具。

五、由本研究所獲得的口試面向、類別、與問題等架構，亦可供作規劃師資培育課程的參考大綱，或作為辦理師資培育的「最後一哩」（last mile）教育之用，以落實教育實習的目的，縮短初任教師適應教師生涯的差距。

六、在未來的研究上，亦可根據本研究對實務應用的涵義，給予口試委員提出最佳面談訓練的有效處方，亦即把面談訓練的重點放在如何設計「口試題目的擬定」及「定錨評量」的問題上。

七、由於口試題目的內容豐富，且概念構圖法又是一個質量並重的方法學，因此，未來的研究也可以朝向編製實用的評量工具著手。例如，「班級經營」類別即有 853 題口試題目可用，其中，有的是具有理論性的問題，有的則是非常實務的問題。因此，可以利用這些題目當成是編製量表的重要題幹，然後，再運用概念構圖及其他測驗編製的技術等方法，去驗證其效度和信度，則一份可用來評量教師「班級經營」能力的量表，即由然而生。至於其他類別問題的研究，亦同。

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

96 年 7 月 15 日

報告人姓名	余民寧(共同主持人)	服務機構及職稱	國立政治大學教育學系教授
時間 會議 地點	96/07/09~96/07/13 日本	本會核定 補助文號	計畫編號：NSC 95-2413-H-004-006
會議 名稱	(中文)心理計量社群國際會議 (英文)The 72nd Annual Meeting of the Psychometric Society		
發表 論文 題目	(中文)層級二單位數目對多層次結構方程式模型有影響嗎？ (英文) Does Number of Level-2 Units in Multilevel Structural Equation Modeling Matter?		

報告內容應包括下列各項：

一、參加會議經過

本次會議與會人員眾多，也有不少來自台灣等地的亞裔人士，在簡單的報到手續後，大家便進入禮堂，大會在主席簡單致辭後，由國際知名的計量大師針對脊迴歸之最新數理公式發展發表專題演講，接著便開始我們的論文發表。在論文發表期間，有多位學者前來索取本人論文之相關資料，並詢問相關的研究細節。此外，個人國科會研究主題如結構化教師甄試口試亦吸引許多學者之關注。許多東西方學者皆相當感興趣，透過面對面的溝通，彼此亦留下聯繫的方式，並開啟日後跨文化比較與合作之契機，實為此行最大的收穫。而本人在會中亦巧遇到多年不見之好朋友，其在國外學術界有驚人的成就。

二、與會心得

本人從國外發表的文章中開拓了研究視野，獲得許多建設性的建議與知識。此外，亦認識了幾位在此領域的國際友人，拓展了人際關係，有助未來的研究順利進行與合作關係。

三、建　　議

未來可以多鼓勵年輕學者至少參加一次國外舉行的國際會議，以開拓視野，激發努力向上的意志力，對過內的學術有一定的提升作用。

四、攜回資料名稱及內容

IMPS2007 會議手冊、名片、當地傳統手工藝品

五、其　　他

無

Does Number of Level-2 Units in Multilevel Structural Equation Modeling Matter?

Ren-Hau Li
Science Education Center, National Taiwan Normal University
davidrhlee@yahoo.com.tw

Min-Ning Yu Yueh-Luen Hu
National Chengchi University, Taiwan

Abstract

How to determine the number of level-2 units in multilevel structural equation modeling (MSEM) as a standard applied to nested or hierarchical data structure was still unknown. This research used Canada data in the large database "Programme for International Student Assessment 2003" (PISA 2003) to check the model-fit indexes and parameters stability in our proposed empirical example processed by MSEM under different numbers of level-2 units. Our proposed example model was first be handled to fit Canada data (26884 students, 948 schools), and then the stabilities of the estimated parameters in the example model under 120, 240, 360, 480, 600, 720, 840 level-2 units were compared. Level-1 units in each school less than 10 students will be crossed out in advance. Besides, intraclass correlations of all variables were controlled in a specified range in different numbers of level-2 units. Finally, we found the ratio of the number of level-2 units relative to the number of estimated parameters of between-level in the multilevel model were 8:1.

Keywords: *Multilevel structural equation modeling, Intraclass correlation,PISA2003*

1. Introduction

“Multilevel” is an important concept in survey data collection and analyses. When research data are collected from hierarchical sampling design, or when nested data structure are obtained due to cluster sampling or multi-stage sampling, traditional statistical analysis methodology would be improper for these data [1-3]. This kind of data derived from clustered or hierarchical sampling designs should be better analyzed by the statistical methods considering data property with clustered, hierarchical or multilevel characteristics. When multilevel characteristics of data are dealt with traditional statistical analysis, the chi-square test of model fit is often inflated, particularly for data with large intraclass correlation (ICC), large group sizes, and highly correlated variables; therefore better fit statistics can not be provided [4-7].

In this study, we would focus on the number of level-2 units in multilevel analysis. When number of between-level groups gradually increased, the inadmissible solutions gradually decreased [8]. Although more level-2 units could be beneficial to obtain admissible solutions and to reduce biases of estimates and standard errors [8], there were no guidelines for us to follow. As a matter of fact, even the appropriate sample size in the traditional structural equation modeling analysis thus far has been inconclusive. An exhaustive examination of the effects on structural equation modeling based on maximum likelihood estimator by Monte Carlo simulation showed that samples fewer than 100 subjects were destructive to ML estimator and larger than 200 subjects were suggested [9]. Tanaka pointed out that there was some agreement on sample-size appropriateness by considering the ratio of the number of subjects to the number of parameters estimated in structural equation modeling with latent variables [10]. Although he did not offer a suggestion about the ratio, he actually explained why the transformation from concerning the ratio of the number of subjects to the number of variables in multiple regression analysis to concerning the ratio of the number of subjects to the number of parameters estimated. Kline indicated that although no absolute standards in the literature of structural equation modeling were offered on the ratio, he suggested the ratio 20:1 be a desirable goal and the ratio of 10:1 be a more realistic target [11].

In regard with the number of level-2 units in multilevel structural equation modeling, in general, though no conclusive suggestion is followed, a larger sample size was usually recommended and preferred [12] that ICC calculated from a small number of groups might not produce reliable estimates and it would be most useful when calculated based on beyond 30 groups. In addition, some studies suggested 50 to 100 groups with at least two individuals nested within each group for multilevel covariance structure modeling [3,4], but the complexity of model was not taken into account in their suggestions. Hence, even groups less than 50 may be enough to get a good model fit. For example, in a study [6] where multilevel confirmatory factor analysis model was used to extract one factor from four measurement items on motivation, and a surprising good model fit of group-level structure based on only 39 groups. However, in another study where two factors with six measurement indexes were modeled in within-level model and only one factor in between-level model, the finding showed that inadmissible estimate problem occurred in the between-level model when group-level sample size was small (50 groups in his research) and ICC was low [8]. Hence, a conclusion was made that the group-level sample size at least 100 would be a better way to deal data with unbalanced groups under Muthén’s

pseudo-balanced solution [8].

In sum, it seemed that the number of level-2 groups at least larger than 30 and the number that did not result in failure in iteration procedures were necessary conditions for multilevel data analysis, but the complexity of model was still not be taken into account to guide reasonable number of level-2 units. Maybe larger level-2 units were good suggestion, nonetheless, just as the chi-square value would be inflated in conventional structural equation model when the number of subjects was too large. Rationally, too large number of level-2 units might also result in inflated chi-square value. Therefore, to formulate an easy regulation for researchers to execute sampling work was an important and meaningful event.

Before checking the number of level-2 units, we had to decide a multilevel model. We adopted "the direct consensus model" based on variable type [15] to extend the same relationships from within-level structure to between-level structure. This was consistent to "homologous multilevel model" [16], which meant that both constructs and functional relations of these constructs in different level were identical. The structure part of the multilevel model was shown in Figure 1, symbol "TEACHSU" represented teacher support perceived from students, "INTMAT" represented interest and enjoyment for learning mathematics, "MATHEFF" represented mathematics self-efficacy, "INSTMOT" represented instrumental motivation to learn mathematics, and "MATH" represented mathematics performance.

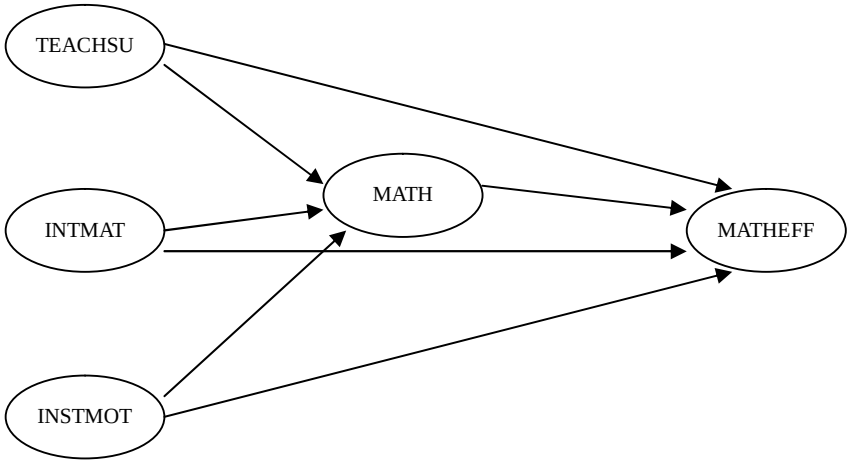

Figure 1. Two-level structural model of the homologous multilevel model.

2. Method

2.1. Subjects

All subjects in this study were obtained from Canada in PISA 2003 [17]. Totally, 26,884 15-year-old students from 948 schools were used in multilevel analysis. We deleted some schools with students fewer than 10 in advance to avoid some outlier cases. Hence, there were at least 11 students in each retrieved school. The reason why we chose Canada data was there were enough level-2 units to process between-level structural equation modeling.

We used all 26,884 subjects from 948 schools for multilevel structural equation modeling analysis to fit the proposed model. Next, we randomly sampled seven different samples from 948 schools with replacement. These seven samples were of different number of schools as 120, 240, 360, 480, 600, 720, 840 schools with 3358, 6959, 10542, 13440, 17160, 20583, and 23900 subjects, respectively.

2.2. Instrument

The measurement indicators were retrieved from student questionnaires in PISA 2003 database. We used all items involving five main factors in my proposed structural equation models, and then exhibited their intraclass correlation (ICC) as shown in Table 1. All analyses were handled with statistical software Mplus 4.0 [18].

Table 1. ICC for each indicator of the five factors

Mathematics		INTMAT		INSTMOT		MATHEFF		TEACHSU	
content	ICC	item	ICC	item	ICC	item	ICC	item	ICC
G1	0.133	I1	0.037	I2	0.033	E1	0.060	T1	0.078
G2	0.171	I3	0.050	I5	0.042	E2	0.062	T2	0.072
G3	0.166	I4	0.041	I7	0.032	E3	0.047	T3	0.074
G4	0.142	I6	0.049	I8	0.035	E4	0.045	T4	0.079
						E5	0.055	T5	0.059
						E6	0.055		
						E7	0.045		
						E8	0.043		

3. Results

The multilevel structural equation modeling analysis was processed simultaneously based on S_{PW} and Σ_B matrixes in respective level. The important parameter estimates were presented in Figure 2. Note that all parameter estimates were under admissible solutions in student-level model and school-level model except that two of four mathematics grades indicators in school-level model were fixed as 0.002 for identification purposes. The degrees of freedom of the multilevel model could be calculated with respective level and then they were combined together. That is, $[25 \times (25+1)/2-60]+[25 \times (25+1)/2-60+2]=265+267=532$.

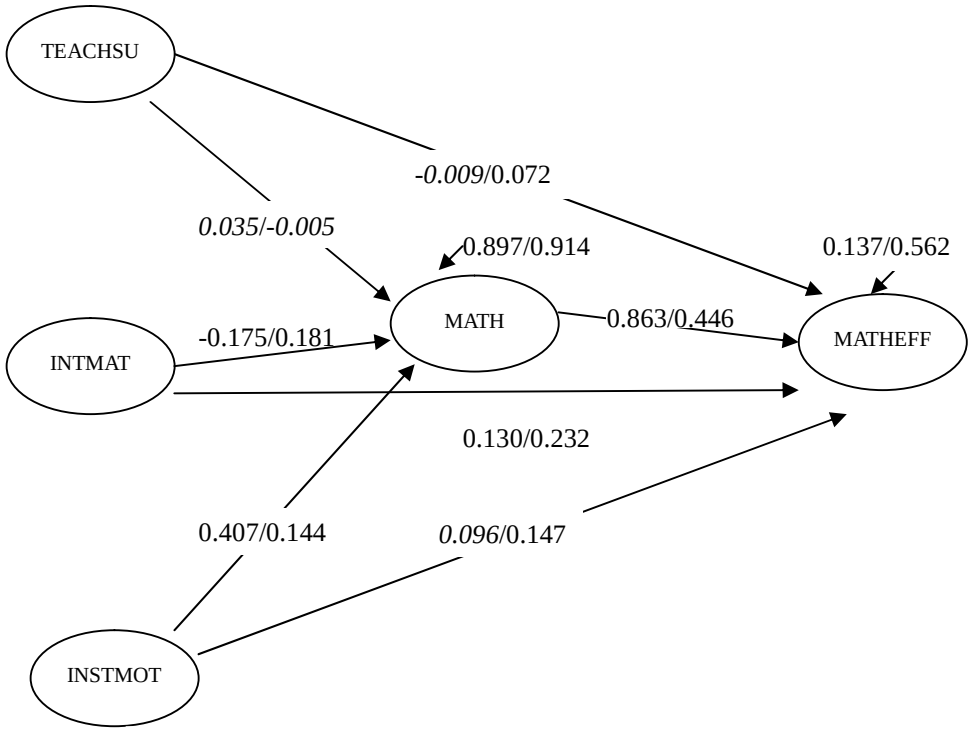

Figure 2. Two level structural model.

Note. Parameter estimates denote: between-level figure/within-level figure

Since the multilevel structural equation model using 948 schools gave a good model fit as shown in Table 2, the effects of level-2 sample sizes were in turn assessed under seven different sets of school-level samples, which were randomly derived with replacement from 948 schools of Canada from PISA 2003 database. These seven sets of samples were arranged to suit the multilevel structural equation model respectively. We would first focus on the comparisons of model-fit indexes in these different level-2 sample sizes, and then the parameter estimates and their estimated standard errors of the multilevel structural equation model were compared in different level-2 sample sizes. In order to have a reference we also listed the results from 948 schools.

Table 2. Model-fit indexes of different school-level sample sizes.

Number of schools	Total sample size	Average school size	Average ICC	$\chi 2$	CFI	TLI	RMSEA	SRMR (Bet./Win.)
120	3358	27.877	0.071	3123.947	0.953	0.948	0.038	0.092/0.035
240	6959	28.935	0.066	6123.788	0.952	0.946	0.039	0.090/0.037
360	10542	29.248	0.064	9184.205	0.952	0.946	0.039	0.084/0.036
480	13440	27.980	0.070	11220.048	0.953	0.947	0.039	0.076/0.036
600	17160	28.581	0.071	14616.772	0.952	0.946	0.039	0.073/0.036
720	20583	28.570	0.067	17378.802	0.952	0.946	0.039	0.075/0.036
840	23900	28.439	0.067	19687.186	0.953	0.947	0.039	0.069/0.036
948	26884	28.346	0.068	22150.368	0.953	0.946	0.039	0.069/0.036

As school-level sample sizes increased from 120 to 948, the total sample sizes inevitably became larger gradually. But the average school size and average ICC of the twenty-five variables were controlled within a limited range. As shown in Table 2, as the school-level sample sizes increased, the average school sizes were almost the same at 28 students or so and the average ICC values were similar with few changes from 0.064 to 0.071. In addition, as expected, the χ^2 values gradually increased as total sample sizes or school-level sample sizes increased. However, the model-fit indexes CFI, TLI, RMSEA and within-level SRMR almost did not change as school-level sample sizes increased; namely, they were almost comparable in different school-level sample sizes. The exception was SRMR model-fit index for the between-level model. As school-level sample sizes increased, the school-level SRMR gradually decreased.

Moreover, although total sample sizes were large, the CFI, TLI, RMSEA and student-level SRMR were almost no change as school-level sample sizes increased, this might suggest that these model-fit indexes were insensitive not only to the different school-level sample sizes but also to the total sample sizes. Besides, since different school-level sample sizes were manipulated and some confounding sources were controlled in the experiment, and the school-level SRMR with obvious changes was the only corresponding outcome, the school-level sample sizes actually determined the degree of goodness-of-fit of between-level model. Hence, only school-level SRMR could help to determine how many level-2 units were sufficient to process the level-2 structural equation modeling under multilevel structural equation modeling condition. Note that if the SRMR below 0.08 was considered to be regular standard guideline, as mentioned earlier, a sample of level-2 units at least about 480 schools was necessary in this study. From the regular standard of the school-level SRMR of 0.08, the ratio of school-level units to the number of parameter estimates would be 480/(58 or 60), equal around to 8 in this study. That is, there is at least eight times or so for the level-2 units relative to the number of parameter estimates. The ratio is close to

the “ten times” regular standard guideline in conventional structural equation modeling as suggested in Kline’s book [11].

Thus far the degree of goodness-of-fit in multilevel structural equation modeling had been checked. In what follows, the parameter estimates and their estimated standard errors were compared in different school-level sample sizes. The stability of parameter estimates was emphasized especially for between-level model part.

In regard with within-level model, the estimates of factor loadings and residual variances, and their estimated standard errors for the within model of multilevel structural equation modeling were presented in Table 3 and Table 4. All parameter estimates in Table 3 and Table 4 were significant. The factor loadings and residual variances in within model were almost the same across the different school-level sample sizes, and their estimated standard errors, as expected, gradually decreased as school-level sample sizes increased. Similar findings also occurred to other parameter estimates in within model, such as structural paths, factor variances, and their estimated standard errors, as shown in Table 5. Although there existed one or two estimates with relatively change across the different school-level sample sizes in Table 5, such as the relationship of MATHEFF regressed on TEACHSU, the significances of the estimates were of consistence. In sum, the stabilities of parameter estimates in within-level model represented that the number of level-2 units actually did not affect level-1 parameter estimates even in the size of 120 schools.

Table 3. Estimates of factor loadings and their standard errors of the within-level model.

	Number of Schools							
	120	240	360	480	600	720	840	948
MATH								
G1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
G2	1.055 (0.010)	1.058 (0.007)	1.060 (0.006)	1.064 (0.005)	1.064 (0.005)	1.063 (0.004)	1.061 (0.004)	1.062 (0.004)
G3	1.033 (0.011)	1.040 (0.007)	1.045 (0.006)	1.042 (0.005)	1.046 (0.005)	1.044 (0.004)	1.040 (0.004)	1.043 (0.004)
G4	1.035 (0.011)	1.042 (0.008)	1.037 (0.006)	1.039 (0.006)	1.041 (0.005)	1.040 (0.005)	1.039 (0.004)	1.040 (0.004)
INTMAT								
I1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
I3	1.146 (0.023)	1.128 (0.015)	1.098 (0.012)	1.106 (0.011)	1.109 (0.010)	1.114 (0.009)	1.110 (0.008)	1.112 (0.008)
I4	1.243 (0.024)	1.246 (0.016)	1.235 (0.013)	1.230 (0.011)	1.222 (0.010)	1.238 (0.009)	1.229 (0.008)	1.231 (0.008)
I6	1.110 (0.023)	1.108 (0.015)	1.116 (0.012)	1.100 (0.011)	1.104 (0.010)	1.112 (0.009)	1.107 (0.008)	1.108 (0.008)
INSTMOT								
I2	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
I5	0.949 (0.018)	0.938 (0.012)	0.943 (0.010)	0.965 (0.009)	0.975 (0.008)	0.959 (0.007)	0.965 (0.007)	0.961 (0.006)
I7	1.101 (0.021)	1.108 (0.014)	1.116 (0.011)	1.123 (0.011)	1.131 (0.009)	1.114 (0.008)	1.120 (0.008)	1.120 (0.007)
I8	0.954 (0.019)	0.957 (0.013)	0.955 (0.011)	0.976 (0.010)	0.976 (0.009)	0.961 (0.008)	0.963 (0.007)	0.964 (0.007)
MATHEFF								
E1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
E2	1.137 (0.036)	1.139 (0.026)	1.121 (0.020)	1.137 (0.018)	1.157 (0.016)	1.141 (0.015)	1.137 (0.014)	1.142 (0.013)
E3	1.316 (0.040)	1.303 (0.028)	1.285 (0.022)	1.305 (0.020)	1.310 (0.017)	1.293 (0.016)	1.297 (0.015)	1.298 (0.014)
E4	1.022 (0.034)	1.039 (0.024)	1.043 (0.019)	1.053 (0.017)	1.052 (0.015)	1.045 (0.014)	1.046 (0.013)	1.046 (0.012)
E5	0.946 (0.032)	0.932 (0.022)	0.983 (0.018)	0.953 (0.016)	0.964 (0.014)	0.948 (0.013)	0.941 (0.012)	0.946 (0.011)
E6	1.202 (0.040)	1.202 (0.028)	1.217 (0.022)	1.224 (0.020)	1.223 (0.018)	1.227 (0.016)	1.239 (0.015)	1.235 (0.014)
E7	1.076 (0.038)	1.081 (0.026)	1.137 (0.021)	1.107 (0.019)	1.107 (0.017)	1.101 (0.015)	1.104 (0.014)	1.104 (0.013)
E8	1.022 (0.037)	1.039 (0.026)	1.040 (0.021)	1.067 (0.019)	1.067 (0.017)	1.051 (0.015)	1.069 (0.015)	1.066 (0.014)
TEACHSU								
T1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
T2	1.074 (0.028)	1.042 (0.019)	1.012 (0.016)	1.006 (0.014)	1.009 (0.012)	1.013 (0.011)	1.013 (0.010)	1.012 (0.010)
T3	1.005 (0.025)	1.004 (0.017)	0.995 (0.015)	0.975 (0.012)	0.981 (0.011)	0.989 (0.010)	0.990 (0.009)	0.985 (0.009)
T4	1.120 (0.030)	1.108 (0.020)	1.107 (0.017)	1.087 (0.015)	1.077 (0.013)	1.091 (0.012)	1.094 (0.011)	1.091 (0.010)
T5	0.977 (0.029)	0.964 (0.020)	0.977 (0.017)	0.958 (0.015)	0.958 (0.013)	0.959 (0.012)	0.957 (0.011)	0.957 (0.010)

[a] They were fixed 1.000 as reference indexes.

Table 4. Estimates of residual variances and their standard errors of the within-level model.

	Number of Schools							
	120	240	360	480	600	720	840	948
G1	0.171	0.166	0.167	0.167	0.169	0.169	0.168	0.169
	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
G2	0.036	0.037	0.035	0.035	0.035	0.035	0.035	0.035
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
G3	0.072	0.071	0.071	0.070	0.069	0.071	0.070	0.070
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
G4	0.092	0.092	0.096	0.096	0.098	0.097	0.096	0.097
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I1	0.268	0.265	0.266	0.263	0.258	0.263	0.260	0.260
	(0.008)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
I3	0.187	0.194	0.211	0.199	0.199	0.200	0.198	0.197
	(0.006)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
I4	0.157	0.150	0.154	0.146	0.148	0.147	0.150	0.149
	(0.006)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
I6	0.202	0.203	0.202	0.208	0.204	0.204	0.204	0.203
	(0.006)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
I2	0.179	0.187	0.186	0.189	0.189	0.181	0.183	0.183
	(0.006)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
I5	0.165	0.167	0.169	0.164	0.163	0.163	0.164	0.164
	(0.005)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
I7	0.227	0.224	0.225	0.226	0.221	0.224	0.226	0.224
	(0.008)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
I8	0.216	0.219	0.224	0.219	0.217	0.216	0.219	0.218
	(0.007)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
E1	0.378	0.372	0.359	0.354	0.353	0.358	0.359	0.357
	(0.010)	(0.007)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)
E2	0.329	0.322	0.323	0.314	0.303	0.314	0.314	0.313
	(0.009)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
E3	0.312	0.308	0.308	0.298	0.295	0.301	0.301	0.300
	(0.010)	(0.007)	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)
E4	0.315	0.315	0.295	0.298	0.295	0.297	0.299	0.298
	(0.009)	(0.006)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
E5	0.300	0.287	0.277	0.277	0.272	0.275	0.272	0.272
	(0.008)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
E6	0.429	0.434	0.416	0.422	0.416	0.418	0.419	0.418
	(0.012)	(0.008)	(0.007)	(0.006)	(0.005)	(0.005)	(0.004)	(0.004)
E7	0.446	0.419	0.395	0.404	0.402	0.400	0.403	0.402
	(0.012)	(0.008)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)
E8	0.472	0.469	0.453	0.459	0.454	0.455	0.453	0.453
	(0.013)	(0.009)	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)
T1	0.361	0.361	0.372	0.368	0.363	0.369	0.371	0.367
	(0.011)	(0.007)	(0.006)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
T2	0.302	0.294	0.307	0.305	0.294	0.299	0.298	0.296
	(0.010)	(0.006)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)
T3	0.215	0.195	0.209	0.198	0.201	0.199	0.195	0.197
	(0.007)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)
T4	0.350	0.341	0.344	0.352	0.347	0.345	0.351	0.350
	(0.011)	(0.007)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)
T5	0.474	0.472	0.468	0.482	0.484	0.477	0.483	0.478
	(0.013)	(0.009)	(0.007)	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)
MATH	0.621	0.643	0.648	0.635	0.638	0.644	0.639	0.638
	(0.019)	(0.014)	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)
MATHEFF	0.124	0.123	0.119	0.118	0.120	0.119	0.116	0.118
	(0.007)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)

Table 5. Estimates of structural paths, factor variances, factor correlations and their standard errors of the within-level model.

	Number of Schools							
	120	240	360	480	600	720	840	948
Structural paths								
MATH on TEACHSU	-0.023[a] (0.028)	-0.020[a] (0.019)	-0.032 (0.016)	-0.015[a] (0.014)	-0.014[a] (0.012)	-0.012[a] (0.012)	-0.014[a] (0.011)	-0.007[a] (0.010)
MATH on INTMAT	0.226 (0.035)	0.211 (0.024)	0.253 (0.020)	0.243 (0.017)	0.238 (0.015)	0.234 (0.014)	0.248 (0.013)	0.242 (0.012)
MATH on INSTMOT	0.183 (0.034)	0.216 (0.023)	0.197 (0.019)	0.188 (0.017)	0.202 (0.015)	0.205 (0.014)	0.187 (0.013)	0.191 (0.012)
MATHEFF on MATH	0.255 (0.011)	0.244 (0.008)	0.252 (0.006)	0.251 (0.005)	0.249 (0.005)	0.246 (0.004)	0.243 (0.004)	0.245 (0.004)
MATHEFF on TEACHSU	0.095 (0.014)	0.073 (0.010)	0.059 (0.008)	0.053 (0.007)	0.054 (0.006)	0.053 (0.006)	0.056 (0.005)	0.055 (0.005)
MATHEFF on INTMAT	0.202 (0.018)	0.188 (0.012)	0.179 (0.010)	0.171 (0.008)	0.174 (0.008)	0.172 (0.007)	0.166 (0.006)	0.170 (0.006)
MATHEFF on INSTMOT	0.089 (0.017)	0.094 (0.012)	0.093 (0.009)	0.109 (0.008)	0.100 (0.007)	0.108 (0.007)	0.110 (0.006)	0.107 (0.006)
Factor variances								
TEACHSU	0.356 (0.016)	0.359 (0.011)	0.349 (0.009)	0.356 (0.008)	0.358 (0.007)	0.351 (0.007)	0.355 (0.006)	0.354 (0.006)
INTMAT	0.367 (0.015)	0.376 (0.010)	0.393 (0.009)	0.388 (0.008)	0.390 (0.007)	0.388 (0.006)	0.394 (0.006)	0.390 (0.005)
INSTMOT	0.384 (0.014)	0.406 (0.010)	0.413 (0.008)	0.388 (0.007)	0.389 (0.006)	0.401 (0.006)	0.397 (0.005)	0.395 (0.005)
factor correlations								
TEACHSU with INTMAT	0.117 (0.008)	0.119 (0.006)	0.124 (0.005)	0.124 (0.004)	0.124 (0.004)	0.125 (0.003)	0.127 (0.003)	0.126 (0.003)
TEACHSU with INSTMOT	0.118 (0.008)	0.127 (0.006)	0.125 (0.005)	0.123 (0.004)	0.119 (0.004)	0.125 (0.003)	0.125 (0.003)	0.124 (0.003)
INTMAT with INSTMOT	0.240 (0.010)	0.250 (0.007)	0.264 (0.006)	0.242 (0.005)	0.247 (0.004)	0.253 (0.004)	0.253 (0.004)	0.251 (0.004)

[a] They were not statistically significant at $p<0.05$; all other parameter estimates were significant.

As for between-level model, the estimates of factor loadings and residual variances, and their estimated standard errors for the between-level model of the multilevel structural equation modeling were listed in Table 6 and Table 7. These estimates in Table 6 and Table 7 were all significantly different from zero at 0.05 significance level and their estimated standard errors, as expected, roughly decreased as school-level sample sizes increased. Note that these parameter estimates were a little unstable when school-level samples were lower than 240 schools, such as factor loading of item I3, I4, I6, I8, and residual variances of item G1, I6, E1, E4, E7, and T3. Besides, it was still several estimates were unstable at the 360 school-level sample size, such as factor loadings of item I3 and I8.

Next, we checked for other parameter estimates in between-level model, such as structural paths, factor variances, and their estimated standard errors in Table 8. These parameter estimates in Table 8 changed dramatically as school-level sample size varied, especially as school-level sample sizes were small. Specifically, most structural paths were quite unstable when school-level samples were fewer than 360 schools, and other parameter estimates were unstable when school-level samples were fewer than 240 schools. Therefore, the number of schools beyond 480 or so would be a good choice to have relatively stable parameter estimates. Take the structural path from INSTMOT to MATH for example, the path coefficient unstably changed from 3.250, 1.758, to 0.672 when corresponding school-level sample sizes were 120, 240, and 360 respectively, and then stably changed from 1.096, 1.025, 0.837, 1.241, to 1.147 when corresponding school-level sample sizes were 480, 600, 720, 840 and 948 respectively. As for all estimated standard errors of these estimates still, as expected, roughly decreased as school-level sample sizes increased.

As a matter of fact, it was not easy to determine the plausible number of level-2 units based on the stabilities of between-level parameter estimates. As found in Table 8, the changes in structural paths were rather irregular. Nonetheless, the number of 480 schools was the best choice for all. Now the ratio of between-level sample sizes to the number of parameter estimates could be calculated as 480/(58 or 60), equal to around 8. That is, at least eight times for the number of level-2 units to the number of parameter estimates was warranted for stable estimates.

In sum, from the outcomes of the model-fit indexes and the stabilities of parameter estimates in multilevel structural equation modeling analysis, the plausible ratio of the number of level-2 units to the number of parameter estimates was 8:1.

Table 6. Estimates of factor loadings and their standard errors of the between-level model.

	Number of Schools							
	120	240	360	480	600	720	840	948
MATH								
G1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
G2	1.170 (0.036)	1.164 (0.025)	1.141 (0.019)	1.148 (0.017)	1.152 (0.014)	1.137 (0.013)	1.144 (0.012)	1.144 (0.011)
G3	1.174 (0.037)	1.161 (0.025)	1.120 (0.019)	1.138 (0.017)	1.140 (0.014)	1.119 (0.013)	1.128 (0.012)	1.130 (0.012)
G4	1.174 (0.037)	1.059 (0.024)	1.049 (0.019)	1.051 (0.017)	1.048 (0.014)	1.040 (0.013)	1.048 (0.012)	1.046 (0.012)
INTMAT								
I1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
I3	0.389 (0.236)	0.638 (0.237)	0.789 (0.131)	0.870 (0.111)	0.959 (0.079)	0.944 (0.080)	0.908 (0.078)	0.939 (0.070)
I4	1.778 (0.307)	1.812 (0.280)	1.350 (0.124)	1.338 (0.108)	1.250 (0.070)	1.213 (0.070)	1.268 (0.070)	1.229 (0.061)
I6	2.142 (0.397)	2.344 (0.400)	1.445 (0.146)	1.517 (0.133)	1.329 (0.081)	1.196 (0.076)	1.286 (0.078)	1.263 (0.069)
INSTMOT								
I2	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
I5	0.991 (0.123)	1.127 (0.122)	1.043 (0.087)	1.048 (0.061)	1.090 (0.054)	1.068 (0.056)	1.180 (0.060)	1.110 (0.050)
I7	0.922 (0.141)	0.971 (0.116)	1.013 (0.086)	1.005 (0.063)	0.995 (0.053)	1.053 (0.057)	1.047 (0.056)	1.027 (0.049)
I8	0.670 (0.124)	0.888 (0.112)	0.873 (0.085)	0.932 (0.062)	0.953 (0.055)	0.946 (0.056)	0.938 (0.054)	0.952 (0.048)
MATHEFF								
E1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
E2	1.016 (0.147)	1.009 (0.104)	1.130 (0.085)	1.080 (0.076)	1.067 (0.062)	1.076 (0.057)	1.044 (0.053)	1.055 (0.051)
E3	1.109 (0.158)	1.028 (0.107)	0.960 (0.072)	0.935 (0.065)	0.942 (0.056)	0.886 (0.051)	0.937 (0.048)	0.938 (0.046)
E4	0.912 (0.141)	0.938 (0.098)	0.828 (0.065)	0.897 (0.063)	0.871 (0.052)	0.882 (0.048)	0.897 (0.044)	0.906 (0.043)
E5	0.892 (0.129)	0.906 (0.094)	0.815 (0.066)	0.815 (0.060)	0.848 (0.053)	0.798 (0.047)	0.826 (0.044)	0.841 (0.042)
E6	0.749 (0.157)	0.645 (0.112)	0.580 (0.082)	0.677 (0.076)	0.733 (0.066)	0.596 (0.059)	0.685 (0.056)	0.677 (0.053)
E7	0.780 (0.137)	0.795 (0.104)	0.604 (0.074)	0.682 (0.068)	0.688 (0.060)	0.600 (0.052)	0.695 (0.050)	0.685 (0.047)
E8	0.523 (0.150)	0.479 (0.107)	0.449 (0.075)	0.526 (0.071)	0.495 (0.060)	0.370 (0.053)	0.523 (0.051)	0.466 (0.048)
TEACHSU								
T1	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]	1.000[a]
T2	1.259 (0.152)	0.960 (0.076)	0.972 (0.066)	0.873 (0.053)	0.890 (0.047)	0.936 (0.046)	0.900 (0.040)	0.903 (0.038)
T3	1.152 (0.127)	0.911 (0.065)	0.923 (0.053)	0.877 (0.044)	0.900 (0.041)	0.921 (0.039)	0.911 (0.035)	0.898 (0.033)
T4	1.309 (0.157)	1.199 (0.086)	1.102 (0.068)	1.182 (0.059)	1.156 (0.053)	1.111 (0.049)	1.095 (0.044)	1.104 (0.041)
T5	0.961 (0.148)	0.905 (0.082)	0.873 (0.064)	0.887 (0.055)	0.885 (0.049)	0.885 (0.047)	0.905 (0.043)	0.904 (0.040)

[a] They were fixed 1.000 as reference indexes.

Table 7. Estimates of residual variances and their standard errors of the between-level model.

	Number of Schools							
	120	240	360	480	600	720	840	948
G1	0.006	0.004	0.002	0.002	0.002	0.002	0.002	0.002
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
G2	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]
G3	0.001	0.000	0.001	0.001	0.001	0.001	0.002	0.002
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
G4	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]	0.002[a]
I1	0.007	0.005	0.005	0.007	0.005	0.004	0.005	0.005
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I3	0.018	0.025	0.020	0.020	0.020	0.020	0.020	0.020
	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
I4	0.002	0.003	0.002	0.002	0.002	0.003	0.002	0.002
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I6	0.002	0.001	0.007	0.005	0.006	0.007	0.006	0.006
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I2	0.001	0.002	0.002	0.002	0.001	0.002	0.002	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I5	0.002	0.003	0.002	0.002	0.003	0.003	0.002	0.003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I7	0.006	0.004	0.002	0.004	0.003	0.003	0.003	0.003
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I8	0.004	0.004	0.004	0.005	0.005	0.005	0.004	0.004
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
E1	0.017	0.014	0.008	0.012	0.009	0.009	0.009	0.010
	(0.004)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
E2	0.009	0.009	0.014	0.015	0.011	0.012	0.012	0.012
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
E3	0.011	0.010	0.007	0.008	0.008	0.009	0.009	0.009
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
E4	0.011	0.007	0.006	0.008	0.006	0.006	0.005	0.005
	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
E5	0.005	0.006	0.008	0.009	0.009	0.009	0.008	0.008
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
E6	0.027	0.026	0.025	0.028	0.025	0.026	0.026	0.025
	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
E7	0.010	0.015	0.018	0.019	0.018	0.016	0.016	0.016
	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
E8	0.028	0.025	0.020	0.025	0.020	0.019	0.021	0.020
	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
T1	0.008	0.008	0.010	0.011	0.013	0.012	0.012	0.012
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
T2	0.010	0.009	0.016	0.016	0.014	0.014	0.013	0.013
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
T3	0.001	0.003	0.004	0.005	0.006	0.005	0.005	0.005
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
T4	0.009	0.006	0.011	0.008	0.008	0.009	0.008	0.008
	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
T5	0.016	0.012	0.012	0.013	0.013	0.012	0.013	0.012
	(0.005)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
MATH	0.103	0.119	0.113	0.115	0.124	0.120	0.114	0.117
	(0.033)	(0.016)	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)
MATHEFF	0.003	0.003	0.002	0.003	0.003	0.004	0.004	0.004
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

[a] They were fixed 0.002 for identification purposes.

Table 8. Estimates of structural paths, factor variances, factor correlations and their standard errors of the between-level model.

	Number of Schools							
	120	240	360	480	600	720	840	948
Structural paths								
MATH on TEACHSU	0.189[a] (0.419)	0.072[a] (0.196)	0.107[a] (0.131)	0.062[a] (0.117)	-0.012[a] (0.110)	0.071[a] (0.097)	0.038[a] (0.090)	0.058[a] (0.085)
MATH on INTMAT	-3.449[a] (2.303)	-2.128[a] (1.183)	-0.146[a] (0.429)	-0.797 (0.390)	-0.367[a] (0.311)	-0.183[a] (0.255)	-0.404[a] (0.239)	-0.501 (0.236)
MATH on INSTMOT	3.250[a] (1.956)	1.758 (0.824)	0.672[a] (0.391)	1.096 (0.290)	1.025 (0.293)	0.837 (0.248)	1.241 (0.243)	1.147 (0.224)
MATHEFF on MATH	0.316 (0.065)	0.364 (0.037)	0.402 (0.028)	0.399 (0.026)	0.381 (0.021)	0.395 (0.020)	0.382 (0.019)	0.385 (0.017)
MATHEFF on TEACHSU	0.130[a] (0.102)	0.093[a] (0.054)	-0.026[a] (0.041)	-0.047[a] (0.037)	0.003[a] (0.032)	0.007[a] (0.031)	-0.007[a] (0.028)	-0.007[a] (0.026)
MATHEFF on INTMAT	-0.057[a] (0.609)	0.067[a] (0.311)	0.344 (0.137)	0.333 (0.126)	0.220 (0.092)	0.097[a] (0.082)	0.193 (0.075)	0.166 (0.072)
MATHEFF on INSTMOT	0.418[a] (0.543)	0.176[a] (0.226)	0.146[a] (0.123)	0.086[a] (0.095)	0.094[a] (0.088)	0.110[a] (0.081)	0.142[a] (0.078)	0.120[a] (0.071)
Factor variances								
TEACHSU	0.028 (0.007)	0.040 (0.006)	0.044 (0.006)	0.048 (0.005)	0.048 (0.005)	0.044 (0.004)	0.046 (0.004)	0.047 (0.004)
INTMAT	0.009 (0.004)	0.005 (0.002)	0.012 (0.003)	0.012 (0.002)	0.018 (0.002)	0.016 (0.002)	0.015 (0.002)	0.016 (0.002)
INSTMOT	0.014 (0.004)	0.011 (0.003)	0.014 (0.003)	0.019 (0.003)	0.019 (0.002)	0.017 (0.002)	0.014 (0.002)	0.016 (0.002)
factor correlations								
TEACHSU with INTMAT	0.008 (0.003)	0.007 (0.002)	0.010 (0.002)	0.012 (0.002)	0.016 (0.002)	0.012 (0.002)	0.012 (0.002)	0.014 (0.002)
TEACHSU with INSTMOT	0.010 (0.004)	0.010 (0.003)	0.010 (0.002)	0.013 (0.002)	0.015 (0.002)	0.012 (0.002)	0.012 (0.002)	0.013 (0.002)
INTMAT with INSTMOT	0.009 (0.003)	0.006 (0.002)	0.009 (0.002)	0.011 (0.002)	0.014 (0.002)	0.012 (0.002)	0.010 (0.001)	0.011 (0.001)

[a] They were not statistically significant at $p<0.05$; all other parameter estimates were significant.

4. Conclusion

In this study where we took model complexity into account and used empirical data from PISA 2003 database instead of simulation data. Besides, we controlled some confounding variables, such as intraclass correlation and school size. We found that the sample size relative to the number of model parameter estimated in between level of structural equation modeling was close to at least 8:1. The finding was consistent to past suggestion on traditional structural equation modeling analysis in literature and would be beneficial to future sampling design in multilevel study.

Reference

[1] Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

[2] Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

[3] Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

[4] Duncan, T. E., Alpert, A., & Duncan, S. C. (1998). Multilevel covariance structure analysis of siblings antisocial behavior. *Structural Equation Modeling, 5,* 211-228.

[5] Farmer, G. L. (2000). Use of multilevel covariance structure analysis to evaluate the multilevel nature of theoretical constructs. *Social Work Research, 24*(3), 180-189.

[6] Li, F., Duncan, T. E., Duncan, S. C., Harmer, P., & Acock, A. (1997). Latent variable modeling of multilevel intrinsic motivation data. *Measurement in Physical Education and Exercise Science, 1*(4), 223-244.

[7] Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22,* 376-398.

[8] Hox, J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8,* 198-207.

[9] Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality.* Unpublished doctoral dissertation, University of Groningen.

[10] Tanaka, J. S. (1987). "How big is big enough ?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134-146.

[11] Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

[12] Bliese, P. D., & Halverson, R. R. (1998). Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management, 24*(2), 157-172.

[13] Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

[14] Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338-354.

[15] Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*(2), 234-246.

[16] Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods, 3,* 211-236.

[17] Organization for Economic Co-operation and Development [OECD] (2005). *PISA 2003 technical report.* Paris: OECD.

[18] Muthén, L. K., & Muthén, B. O. (2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.