管 理 與 系 統
第十六卷 第二期 民國九十八年四月
181-200 頁

Journal of Management & Systems
Vol. 16, No. 2, April 2009
pp. 181-200

專家區間估計中過度自信現象之研究

Expert Overconfidence in Probability Interval Estimations

林希偉 Shi-Woei Lin　　**王國全** Kuo-Chuan Wang

元智大學企業管理學系

Department of Business Administration, Yuan Ze University

(Received September 29, 2007; Final Version February 19, 2008)

摘要：決策或風險分析中關於不確定性的評估往往需要專家意見的投入，然而過度自信的專家判斷卻可能影響決策的品質。本研究探討專家機率區間估計中過度自信差異的來源，同時，爲了修正二元性校準衡量的不穩定問題，我們運用專家主觀機率之期望絕對離差與經由實現值所觀測到的絕對離差來定義新的連續型校準指標，並以線性混合模型來分析與詮釋資料。我們發現專家之間的變異小於問題之間或者真實值所造成的隨機變異。因此，實務上使用專家機率區間判斷時，運用種子問題來篩選較具專業知識或校準程度較高的專家可得到的效果恐將有限。

關鍵詞：過度自信、校準、區間估計、專家判斷

Abstract: Expert judgment has been widely applied in the field of decision making and risk assessment. However, when overconfidence reveals in the judgment, it might serious affect decision quality. Thus, this study aims at discussing different sources of overconfidence in an expert's probability interval estimation. To revise the instability which is caused by using binary variables as calibration measuring in previous research, we define a continuous variable as new calibration measurement by applying the ratio between EAD (Expected Absolute Deviation) of experts' subjective probability and MAD (Mean Absolute Deviation) of realization; further, analyze and interpret the data by a simpler linear mixed model. Under the new calibration measurement, we discover that the

* 本論文承兩位匿名審查委員細心審閱並提供諸多寶貴意見，特此致謝。本研究亦感謝國科會補助部分研究經費（NSC 96-2416-H-155-008）。。

variance among experts is less than the random variance among questions or realizations. This result has overthrown the analytic outcome of binary calibration measurement. Thus, to use expert judgment in practice, the effect may be limited by adopting seed questions to select a more professional expert or the one in higher calibration level.

Key words: Overconfident, Calibration, Interval Estimate, Expert Judgment

1. 研究背景與動機

決策分析的特點之一在於它能夠被應用於解決現實生活中包含高度不確定性的複雜問題，取得一個沒有偏誤、可以真實反應出不確定情況的機率估計遂成爲分析工作施行時的重要關鍵。由於決策分析通常被應用於處理不重複的決策（one-time decision），因此，機率估計往往非由資料、而是透過訪談專家來取得，專家判斷（expert judgment）亦因此被廣爲應用在決策或風險分析中。

簡言之，專家判斷是依賴一位或多位專家的經驗來對重要議題進行判斷與評估的方法。在沒有足夠的歷史資料可供參考，或者取得資料的成本過高時，我們往往轉而求助於專家的意見。而由於它具有高時效、低成本的特性，又能夠提供重要資訊以供預測與決策之用，因此亦被廣爲應用在經濟、氣象、軍事、政策分析等領域。

因爲擷取方式與使用目的的差異，專家的機率判斷往往也有不同形式的呈現，我們除了要求專家評估一事件之某特性（即一個隨機變數）最可能出現的數值外，有時亦要求專家提供此變數的機率區間（probability interval）。這些區間估計可以運用在不同的決策分析模型中，例如包含不確定機率（imprecise probability）或不確定效用函數的決策樹，或者是貝氏統計推論中先驗機率的建構。由於區間估計往往比點估計提供更多的訊息，亦可能提昇決策的品質，因此不只受到實務使用者的偏愛，近年來也獲得研究人員的高度重視（Baginski *et al*., 1993; Chatfield, 2001; Johnson, 1982; Walley, 1991）。

雖然從主觀機率論者如 de Finetti （1974）的觀點看來，機率全然是個人的內在信念，沒有所謂的「正確」或者「客觀」，比較不同主觀機率（信念）之間的不同亦沒有實質的意義。但是，當專家的意見被當成決策的投入時，我們仍可透過一些機制來檢驗判斷的有效性，舉例來說，若專家提供的是一系列的 90%機率區間，則我們通常可以預期在長期觀察下，專家提供的所有區間估計中，應有 90%會包含各個變數所對應的實現值（realization）。但根據 Russo and Schoemaker （1992）的研究，當專業經理人被要求對相關領域的重要議題提供 90%信賴區間時，只有約 40%~60%的區間會包含實現值。而在其他如 Cooke（1991）和 Shlyakhter *et al*.（1994）的研究中，亦指出專家的區間估計通常存在過於狹窄的問題。

在專家判斷的研究中，和上述發現有關，而且也最廣爲探討的議題就是專家的校準（calibration）。校準是主觀機率判斷之精準度的衡量，它原本是認知心理學中用來討論人類關於機率的主觀認知與理想狀況之間的接近程度，因此，若專家評估一系列獨立事件，且認定這些事件發生的主觀機率均爲 P，而我們又可以在事後透過觀察而得知事件之發生與否，則對一位校準良好的專家而言，我們將期待這些被評估事件的相對發生頻率和專家的主觀機率 P 相去不遠。然而對於使用專家判斷的決策者而言，校準卻未必是唯一的考量，因爲較窄的區間可能代表更高的資訊價值，因此校準和明確度（sharpness）就成爲重要但又可能相互衝突的兩個標準，如果不能同時達到一定水準，則該區間估計的實用性可能有限。然而，這也可能促成專家提供過於狹窄的機率分佈，即大部分的真實值都出現在預期分佈（例如 90%的區間）以外的極端部份。我們稱這種現象爲過度自信（overconfidence）。

早期研究顯示一般受測者在回答常識問題時常會出現過度自信現象（Lichtenstein and Fischhoff, 1977），Kahneman and Riepe（1998）亦指出人們常常過於樂觀，低估了風險並且高估了自己解決問題的能力，Plous（1993）除了指出多數人存有過度自信的心理現象，他甚至認爲「過度自信應該是在決策心理相關議題中最禁得起考驗的發現」。而過度自信也非僅存在於一般人的決策判斷之中，許多研究證實了專家的判斷往往也存有過度自信（如 Morgan and Henrion, 1990; Russo and Schoemaker, 1992; Shlyakhter *et al*., 1994）。近年來關於校準與過度自信的研究則著重在探討過度自信現象究竟是一種行爲偏誤，或者是源自於區間估計擷取形式的影響（Juslin *et al*., 1999; Klayman *et al*., 1999； Soll and Klayman, 2004）。其他的研究則著重在探討這些偏誤是否來自於問題選擇時的偏差，因若將難度不同的問題放在一起比較，困難的問題往往較容易導致過度自信的產生（Budescu *et al*., 1997a, b; Juslin *et al*., 2000; Klayman *et al*., 1999）。然而，不管專家發生過度自信之原因爲何，過度自信終將使所得之區間估計過度狹窄，而導致錯誤的決策。

再者，過去的研究中，關於專家判斷之校準的探討並不多，亦不夠完整。主要原因在於過去實證研究中所採用的資料多是以研究生或者一般社會大眾爲研究對象，而他們在實驗過程中所回答的問題也多是一般常識問題，因此所得之結論與現實各領域中應用專家判斷的情形可能存有相當落差。Lin and Bier（2008）試圖透過分析一個由真實世界中各領域專家所提供的區間估計集合而成的資料庫，來更明確地探討影響校準的因素。他們的模型對造成專家過度自信的關鍵因素提出初步的看法，但分析中仍有未臻完美之處。其中最大的問題在於他們使用了二元計分法，亦即根據一個個別區間估計是否包含實現值來予以 1 或 0 的校準分數。然而，由於專家的區間估計與其過度自信的程度本屬連續型的衡量，因此在轉換爲二元計分的過程中，可能加大校準計分的變異程度。舉例來說，如果二位專家提供了相似的區間估計，但其中一位卻具有「些微」過度自信的傾向，則相較於得到校準分數爲 1 的另一位專家，他可能只因爲提供略

爲窄小的區間估計而得到分數爲 0 的校準度。因此，如果我們沒有大量的問題可以用來求算專家的校準平均，使用具較大的變異性的二元因變數將可能造成分析模型的效度問題；再者，二元因變數在分析時需要使用廣義線性混合模型（generalized linear mixed model），求算其最大概似估計時的高維度積分問題往往造成運算的困難，然而當使用限制擬似概似估計法（panelized quasi-likelihood）或貝氏估計法來取代最大概似估計時，又可能造成模型參數估計的偏誤。此外，該廣義線性混合模型中的諸多假設不只尚未經過檢測，嚴謹的檢測方法與程序也還有待統計學家與應用數學家進一步研究開發。

因此，我們在本研究中延伸 Lin and Bier（2008）的研究，對專家的校準與過度自信提供一個實證的分析。在這個分析中，區間估計的資料是由一群各領域之中具相關知識經驗的專家，針對該領域的重要實際議題所給予的回應，並非由學生或外行人回答一般性常識問題所得。同時，我們亦藉由使用不同的分析方法與採用不同的校準衡量來對過度自信議題做更進一步的探討。我們以 Soll and Klayman（2004）的校準衡量爲基礎，但改以期望絕對離差（expected absolute deviation, EAD）與絕對離差（absolute deviation, AD）發展出一個連續型校準衡量來取代之前的二元回應變數，這個衡量修正了二元回應變數不穩定的問題。我們同時改採線性混合模型來分析專家在新校準衡量下的過度自信現象，並與使用二元回應變數所得的分析結果相互比較。

本研究的主要目的並不在於了解造成過度自信的潛在因子，雖然這類研究不論在實務或者理論上均有其重要性，但是往往需要在嚴謹控制的實驗環境中進行。反之，本研究希望能夠著重在決策分析實務上使用專家區間估計時會考量到的幾個重要議題。例如，對於是否應該試圖去擴展專家所提供的信賴區間或者機率分佈，學者之間往往有不同的意見，因此，了解不同專家之間過度自信的程度與其變異大小將有助於我們得知是否單純使用數學整合方法來彙整專家意見就可以達到良好的校準，或者我們必須使用更複雜的模型來加寬專家提供的區間（舉例來說，如果專家之間的差異不多，則使用傳統加權平均來整合專家判斷的作法可能成效有限）。

接下來，我們將在第二節中簡介研究資料，第三節我們介紹 Soll and Klayman (2004) 的校準衡量，並且呈現專家資料在此校準衡量下的實證分析結果，在第四節中，我們以 Soll and Klayman (2004) 的校準衡量爲基礎，精煉並且發展出新的校準衡量，我們接著以線性混合模型進行統計分析，並與過去相關研究結果比較。最後則爲結論、研究限制與後續研究的建議。

2. 研究資料簡介

本研究採用的專家判斷資料是由荷蘭學者 Cooke 與其他歐洲的風險分析專家進行的一系列專家判斷相關研究中所取得的數據。這其中包括了 27 個不同的專家小組，他們各自對不同實務上需要解決的問題提出機率區間估計。這些專家小組的專業領域與他們面對的問題均有極大的

差異，除了比較偏向日常商業活動的研究，如股票選擇權的交易風險、荷蘭不同城市房地產租金的預測等，也包括了大量不同產業複雜的風險分析問題，例如合成材料的安全性分析、洪水高度的預測、放射性物質的擴散模型、與輸油管線的安全評估等。這些研究中，或者由於實際數據無法取得，或者由於數據取得的相對成本過高，因而改採專家提供的意見做爲決策的基礎。由於篇幅的限制，我們僅在表 1 列出個別研究的代號與簡單說明，希望得到詳細參考文獻資料的讀者，請參考 Lin and Bier（2008）。

在不同的專家判斷研究中，對於每個問題，小組中的每個專家必須提供至少 3 個分位數（即第 5、第 50 與第 95 分位數）的估計。而在少數的幾個研究中，專家還被另外要求提供第 25 和第 75 分位數。雖然在不同的研究中，擷取專家意見的操作方法可能稍有差異，而且少數甚至是使用圖形而非數字的型式來完成，但原則上，在說明研究的目的與進行機率估計的練習之後，實際操作的擷取過程可能只是簡單地要求專家「提供某個不確定數量（例如阿姆斯特丹的頂級辦公大樓在 1998 年第一季的租金）的第 5、25、50、75 與 95 分位數」。

在這些研究中，專家被要求回答的問題包括種子問題（seed question）以及研究中決策者真正想要得知正確答案的「真實」問題。種子問題是指一些可以用來測試與評估專家的專業知識，和真實問題屬於同一知識領域，難度相當，但實現值（realization）均爲已知的問題。因此這些種子問題的選擇都盡量和真正期待專家去提供預測（但無法得到其實現值）的真實問題具有高度相關。此研究中的數據含蓋了 203 位專家、519 個種子問題以及 4562 個 90%的信賴區間估計。如上所述，這 203 位專家亦回答了大量的真實問題，但由於我們無法取得這些問題的實現值，因此它們的分位數估計與機率區間並沒有被納入分析之中。

3. 連續型校準衡量的初步選擇與分析

二元校準衡量的不穩定，除了可能造成模型參數估計時的收斂問題，同時，它也造成我們對模型與資料的配適程度的質疑，並且進一步質疑模型的穩定性與有效性，因此，我們勢必得發展出一個較佳的方法來衡量專家的校準。爲了徹底解決二元校準衡量的不穩定問題，我們在本研究中採用不同的連續型校準測量來進行分析，並藉以確認之前經由廣義線性混合模型（generalized linear mixed model，以下簡稱 GLMM）所獲得的結論是否依然穩健可靠。原則上，如果專家判斷的衡量尺度爲連續而非二元，則將允許我們使用較簡單且有效的線性混合模型（linear mixed model，以下簡稱 LMM），而不需使用模型參數估計與檢定方法之發展都尚待完成的 GLMM。使用 LMM 將不僅讓我們能夠在模型的參數估計上採用更可靠的最大概似估計法，並可以提供一套模型檢測和診斷的完整工具（例如適合度檢測）。

表 1 專家判斷研究中的專家小組

專家小組代號	研究主題	專家數	種子數
Acrylonitrile	丙烯的劑量反應關係	7	10
Ammonia	氨的劑量反應關係	6	10
Sulphur trioxide	三氧化硫的劑量反應關係	4	7
Option trading	選擇權交易市場中金融產品價格的預測	9	38
Real estate	歐洲城市房地產的租金預測	4	31
Building temperature	建築物的熱循環與舒適度評估	6	48
Dike ring	環形築堤的安全性評估	17	47
Movable barriers	移動式水閘的可靠度評估	8	14
River dredging	水道的疏浚	6	8
Crane risk (flanges)	起重機凸緣連結部分的風險分析	10	8
Crane risk	起重機的風險	8	12
Rocket propulsion	火箭推進系統的風險分析	4	13
Space debris	太空漂浮物的風險分析	7	26
Composite materials	合成材料的安全性分析	6	12
Atmospheric dispersion	放射性物質在大氣中的散佈	8	23
Atmospheric dispersion (TNO)	放射性物質在大氣中的散佈（由荷蘭應用科學研究組織進行的先導研究）	7	36
Atmospheric dispersion (Delft)	放射性物質在大氣中的散佈（由 Delft 科技大學進行的先導研究）	11	36
Radiation dosimetry	放射性物質的體內劑量評估	8	55
Early health effects	放射性物質對健康初期影響的評估	9	15
Wet deposition	放射性物質的濕沈降情形	7	19
Dry deposition	放射性物質的乾沈降情形	8	14
Radioactive disposition (Delft)	放射性物質的沈降（先導研究）	4	24
Radiation in food	食物鍊中經由動物傳遞的放射性物質風險	7	8
Soil transfer	食物鍊中經由植物與土壤傳遞的放射性物質風險	4	31
Groundwater	地下輸水道的風險分析	7	10
Gas pipelines	地下輸油管線的故障頻率	15	28
Montserrat	Montserrat 火山的風險分析	11	8

3.1 連續型校準計分法的選擇

Cooke（1991）以古典假設檢定爲基礎推導出卡方校準計分法—它是基於專家分位數估計所切割出的區間中實現值的預期次數與觀測次數之間的差距所導出的卡方統計量。卡方統計量雖是連續型的校準計分，但此計分法乃是基於統計理論上的大樣本性質（即極限分配），因此需要每個區間裡的觀察值數目夠大（常用的經驗法則是每組皆需要 5 個以上的期望觀測值）。在專家判斷的研究中，由於建構種子問題的困難度很高，所以研究中專家回答的種子問題數目往往不多，因此並不十分適合此計分法，而 Wiper *et al*.（1994）所推導出的 Kolmogorov-Smirnov 計分法則要求專家提供一個完整分配的評估，所以在實際操作上亦有其困難。

除了大樣本性質無法在有限的種子問題中得到良好的逼近，Hora（2004）還認爲上述基於假設檢定推導出的校準計分法僅使用了蘊藏在資料中的部分資訊，因此他建議研究者使用經驗累積分配函數（empirical cumulative distribution function）與累積分配函數之間的差距來衡量校準的高低。然而，Hora（2004）的校準可能需要比我們的資料中更多的分位數估計。因此我們使用 Soll and Klayman（2004）提出的另一種測量校準與過度自信的方法，以下我們將詳細說明這個校準衡量的計算方法。

首先，Soll and Klayman（2004）假設專家對於每個問題的判斷都有其主觀機率密度函數（subjective probability density function, SPDF），他們認爲專家的每個判斷都來自腦中的主觀機率密度函數，所以他們會認爲某些（或某區段的）數值出現的可能性會比其他更高。在本研究中，我們假定每位專家的主觀機率密度函數可以經由分位數的方式被抽離出來。所以，我們亦可以藉由專家提供的第 5、第 50 以及第 95 分位數，來反推出一個適切的機率密度函數。若是專家給予的第 5 與第 95 百分位數相較於中位數而言約略呈對稱，則 Soll and Klayman（2004）已經證實三角或者常態分配等不同的對稱機率密度函數並不會造成分析結果出現太大差異，若分位數指出專家的主觀機率是偏斜的，我們也許應當考慮使用廣義的 Beta 分配。因此，由分位數求得機率密度函數與相關的傳統參數後，我們可透過統計軟體或者風險分析軟體（如@Risk）計算，得到平均期望絕對離差（mean expected absolute deviation, MEAD）以及平均絕對離差（mean absolute deviation, MAD），並且可以計算出下列比例：

$$M = \frac{MEAD}{MAD} \tag{1}$$

由於平均絕對離差（MAD）代表真實值在平均值周圍的平均距離，因此代表理想的區間寬度的衡量，而平均期望絕對離差（MEAD）則代表這些專家所提供的主觀機率中，與平均值的平均距離。因此比例 M 代表的即是實際區間（專家估計的區間）與理想區間大小的比例。在本研究中，由於資料並非高度偏斜，我們假設專家的主觀機率密度函數接近常態分佈，根據常態

分配的特性，我們可以得到期望絕對離差：

$$EAD = \frac{W_i}{Z_{0.5+(p/2)}\sqrt{2\pi}} \tag{2}$$

此處 W_i 是由專家所提供的第 5 和第 95 兩個分位數所形成區間的寬度，p 是該區間所包含的機率，將所有的期望絕對離差加總後除以區間個數，我們可以得到平均絕對期望離差（MEAD）：

$$MEAD = \sum_{i=1}^{N} \frac{W_i}{Z_{0.5+(p/2)}\sqrt{2\pi}}\left(\frac{1}{N}\right) \tag{3}$$

此處 N 是專家所提供之區間估計的個數，Z 則是標準常態分配值。而平均絕對離差（MAD），則是利用上述常態的平均數（區間的中點）與其相對應問題的實現值之差距，取絕對值之後加總並予以平均。因此我們便可以得到一個理想校準的平均長度（MAD）：

$$MAD = \frac{\sum_{i=1}^{N}|D_i|}{N} \tag{4}$$

此處 D_i 是專家對第 i 個變數所提出之區間估計的中點與該變數的實現值之間的距離。將這兩者結合後，我們可以得到一個比例：

$$M = \frac{MEAD}{MAD} = \frac{\sum_{i=1}^{N} W_i}{\left(Z_{0.5+(p/2)}\sqrt{2\pi}\right)\sum_{i=1}^{N}|D_i|} \tag{5}$$

比例 M 所代表的即是「觀察所得的平均區間長度 / 理想校準下的平均區間長度」。

3.2 Soll 與 Klayman 校準衡量下的初步分析結果

我們將 27 個專家小組的研究資料，使用 Soll and Klayman（2004）的 M 值校準計分法進行計算後，呈現於圖 1。

平均期望絕對離差（MEAD）和平均絕對離差（MAD）兩者的比例 M 值代表專家過度自信的程度，這個比例越低，表示專家估計的區間和理想區間的比值越小，所以呈現出過度自信，反過來，若比值遠大於 1，則表示專家自信不足，因此提供了過寬的區間估計。從圖中我們發現不同的專家小組（即不同主題的專家判斷研究）之平均 M 值校準分數分佈在 0.15 到 1.04 之間，

不同專家小組之間過度自信情形的變異亦相當大。若我們主觀地以 0.8 做爲一個基準，則少數專家小組的表現—例如建築物的熱循環與舒適度評估（代號 Building temperature）、合成材料的安全性分析（代號 Composite materials）、選擇權交易市場中金融商品價格的預測（代號 Option trading）、食物鍊中經由動物傳遞的放射性物質風險（代號 Radiation in food）與歐洲城市房地產的租金的預測（代號 Real estate）—與理想區間寬度已相去不遠。其中合成材料安全性分析的專家小組甚至呈現些微自信不足（underconfidence）的現象。過去專家判斷研究指出，校準度與績效上的回饋可以適度修正專家機率估計中過度自信的問題，也因此某些領域（如氣象預測）的專家在機率估計的校準表現往往優於其他領域的專家。雖然透過跨領域專家小組的資料來更清楚釐清回饋機制對區間估計校準的影響有其重要性，但因爲我們的資料無法對回饋的強度提供適切的衡量，因此我們並不在此深入討論這個議題。

整體而言，我們發現 *M* 值校準分數在多數的研究中低於理想值 1，少數的研究甚至只有 0.2 左右的校準分數，表示其理想的區間寬度幾爲專家平均認知的 5 倍。事實上，27 個專家小組的整體平均 *M* 值亦僅爲 0.54，顯示出專家明顯的過度自信現象。

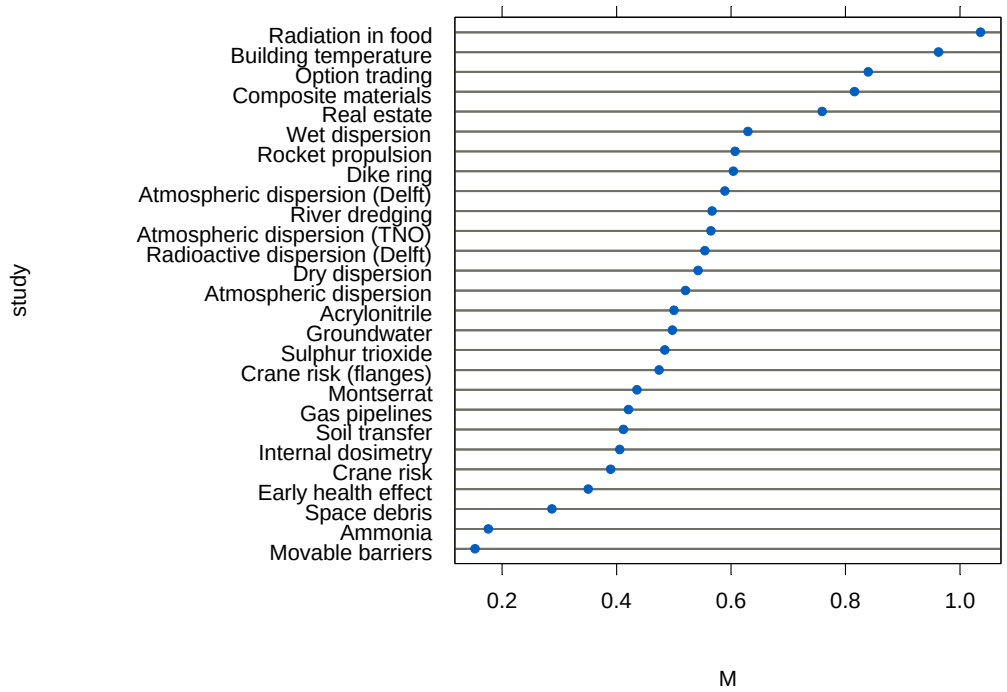

圖 1　不同專家小組在 Soll and Klayman (2004) 的校準衡量下的總體表現

4. 校準衡量的精煉與分析

雖然 Soll and Klayman（2004）的 *M* 值校準衡量可以提供另一個衡量過度自信的指標，且應用在我們的資料庫時，它可以提供每個研究的 *M* 值、提供每個專家（經由回答一系列問題所計算出）的 *M* 值，甚或提供每個問題（基於一個專家小組內的多位專家的意見所計算出）的 *M* 值，但也因爲這個校準衡量是基於二個平均衡量之間的比率，因此我們沒有辦法衡量某個「專家」在回答某個「問題」時的校準，也無法進一步分析專家校準的變異中，那些可以歸因於研究之間的特性，那些屬於專家之間的差異，而那些又來自於問題難易的不同。

同時，由於計算 *M* 值的過程中，我們必須針對不同問題或不同專家的判斷先分別計算其期望絕對離差（EAD）或者絕對離差（AD）的平均值，這個平均後再取比率的動作將會使得 *M* 校準衡量嚴重受到不同問題之原始量尺的影響。因爲平均數本身並不是一個穩健統計量（robust statistic），數值較大的個別 EAD 或者 AD 值將會在平均的過程中主導了 MEAD 與 MAD 的大小，因此，除非我們對所有的變數均採用相似的尺規，而且所得出的個別數值大小也相差不大，否則如果一個專家判斷研究中有部分問題的 EAD 或者 AD 比其他問題小了許多，則不管專家在這些問題的判斷是過度自信或者自信不足，它們的影響都難以顯示在最後整合起來的 *M* 校準衡量之上。

同理可知，這樣的量尺，也可能受到單一極端值的影響，如果其中有一問題的實現值是一個特異值，或者研究人員在紀錄某一問題時因誤用了不同尺度而使得實現值的大小變成其它問題的數萬倍，就可能造成 *M* 值衡量有若大的偏差。基於上述這些理由，我們決定以 Soll and Klayman（2004）的 *M* 值校準衡量爲基礎，發展出一個更具分析上的效度、也更適合我們研究資料的衡量尺度。

4.1 校準計分法的精煉

在我們的新衡量中，我們改採期望絕對離差（expected absolute deviation, EAD）和絕對離差（absolute deviation, AD）做爲基礎。我們定義

$$m = \frac{AD}{EAD} = \frac{\left(Z_{0.5+(p/2)}\sqrt{2\pi}\right)\left|D_i\right|}{W_i} \tag{6}$$

在這個衡量中，我們改採 EAD 爲分母，AD 爲分子，主要原因在於 EAD 的計算乃是基於專家主觀機率密度函數的估計，是由整個分配導出，因此不只數值不會爲零，也相對較爲穩定，而 AD 乃是基於單一實現值，因此變異性較大，亦有可能爲零。所以在這個新的尺規中，分母是專家估計的寬度，分子是理想的寬度，較大的 m 值代表過度自信，而低於 1 的 m 值則代表自

信不足。同時，由於 m 不再是基於二個平均衡量之間的比率，因此可以直接衡量某個「專家」在回答某個「問題」時的校準。

然而，在校準研究的應用上，目前我們所設計的尺規仍有一個問題尚待克服。由於在意見擷取的實務操作中，專家小組通常會由複數專家所組成，因此，不僅個別的專家會被要求回答與其專業知識有關的一系列區間估計問題，針對個別問題，我們也會得到同領域的一群專家所提供的不同估計。因此，當我們計算個別專家（或者個別問題）的校準值時，我們理應計算該專家回答這一系列問題（或者多位專家在回答該特定問題）的數個校準值（m）的平均。但是由於 m 是一個比例的衡量，若欲求其平均值，使用如下的幾何平均應是較合理的作法：

$$\overline{m} = \prod_{i=1}^{n} \left(\frac{AD_i}{EAD_i} \right)^{\frac{1}{n}} \tag{7}$$

然則，幾何平均在統計的分析上卻有其不便利性，所有的線性模型，不論是一般或廣義、固定或混合效應，都是基於算術平均值之計算與檢定。我們若對（7）式取對數，可以得到：

$$m^* = \log \left[\prod_{i=1}^{n} \left(\frac{AD_i}{EAD_i} \right)^{\frac{1}{n}} \right] = \left(\frac{1}{n} \right) \sum_{i=1}^{n} \log \left(\frac{AD_i}{EAD_i} \right) \tag{8}$$

因此，透過對數的轉換，我們不僅可以讓校準衡量的分配更對稱，避開校準衡量 m 容易出現右偏之問題，同時也可以符合統計分析上的運算需求，我們把這個對數轉換過的校準衡量稱爲 m^*。

接下來，我們首先以圖形呈現專家意見在新的校準衡量下之表現，之後，我們把所有的資料整合在一起，使用線性混合模型（LMM）進行整體的分析。

4.2 新校準衡量的初步圖形分析

我們將每組研究資料中使用自然對數轉換後的 m 值之分佈表示如圖 2。從這張盒鬚圖中，我們可以清楚看到各組研究資料的外圍值、第三四分位數、中位數、第一四分位數、以及界外值（outliers）。我們發現，透過對數轉換，不同研究中個別區間估計的校準度分佈看起來大致對稱（雖然仍有部份的研究是屬於偏斜的分佈，但其偏斜情形相對而言並不嚴重）。姑且不論大部分基於變異數分析原理的統計方法都建構在常態假設上，我們知道相關線性模型的分析的過程中，適度的對稱性是最基本的要求，因此，這張圖驗證了對數轉換的合理性。

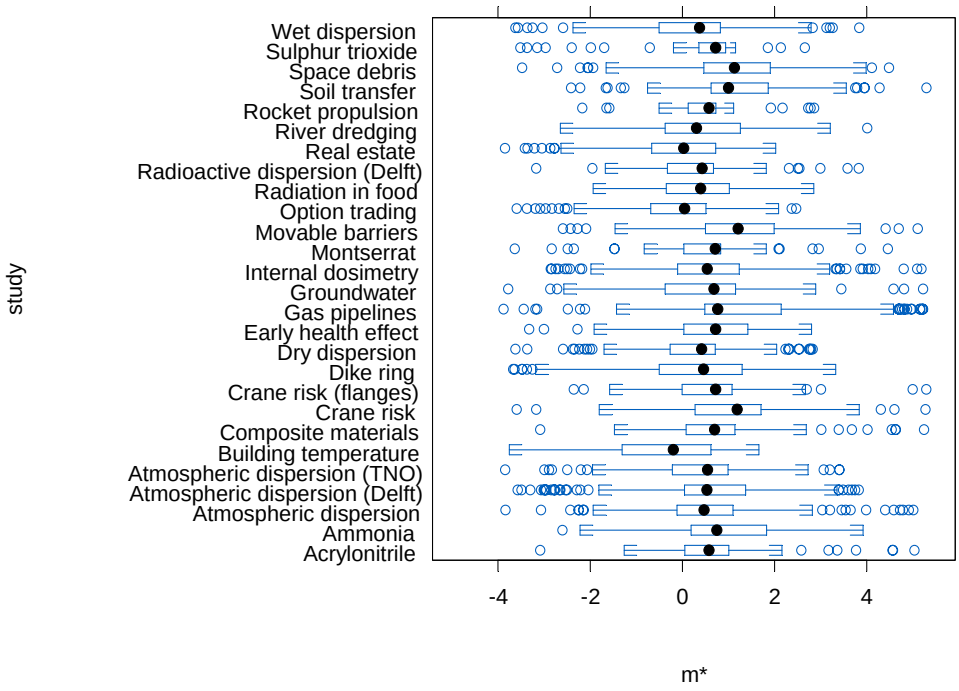

圖 2　不同專家小組區間估計的 *m值分布**

接著我們將各專家小組研究資料中的 *m** 校準分數予以平均，得到如圖 3 之散佈圖。從圖中，我們可以清楚看出對於不同的研究（即不同領域的專家小組），*m* 值平均校準分數在使用對數轉換後仍有很大的變異，其分佈介在-0.2 到 1.2 之間。由於經過對數轉換，所以 *m** 校準的理想值是 0，平均值低於 0 的研究小組代表存在些微自信不足的情況，接近 0 的小組則擁有較佳的校準，代表這些研究中的專家所提供的估計區間很接近理想中 90%區間所應有的寬度。然而，和使用 Soll and Klayman（2004）所建議的 *M* 相同，大部份的專家小組的平均 *m** 校準皆大於 0，代表多數研究中的專家有過度自信的情形發生。

由於我們希望進一步了解校準變異究竟是源自於專家之間的差異（例如源自於專業能力或知識上的不同、機率估計能力的落差，甚或在校準與明確度之間的權衡），或者是來自於問題之間的難易不同。因此，我們在圖 4 中呈現了不同研究中個別專家的 *m** 校準分數，圖中指出在同一個研究中，不同專家間的平均校準差異頗大，但是在不同的研究（專家小組）之間，專家校準分數的變異程度則似乎沒有太大的落差，基本上，並沒有出現某些小組中專家的表現很一致地達到完美校準，而其餘小組就出現參差不齊或變異很大的現象。另外，我們亦發現，雖然以理想值 0 作爲分界時，多數專家的判斷呈現過度自信，但是對分屬不同領域的每一研究來說，往往都有少數的專家擁有接近完美校準的表現。

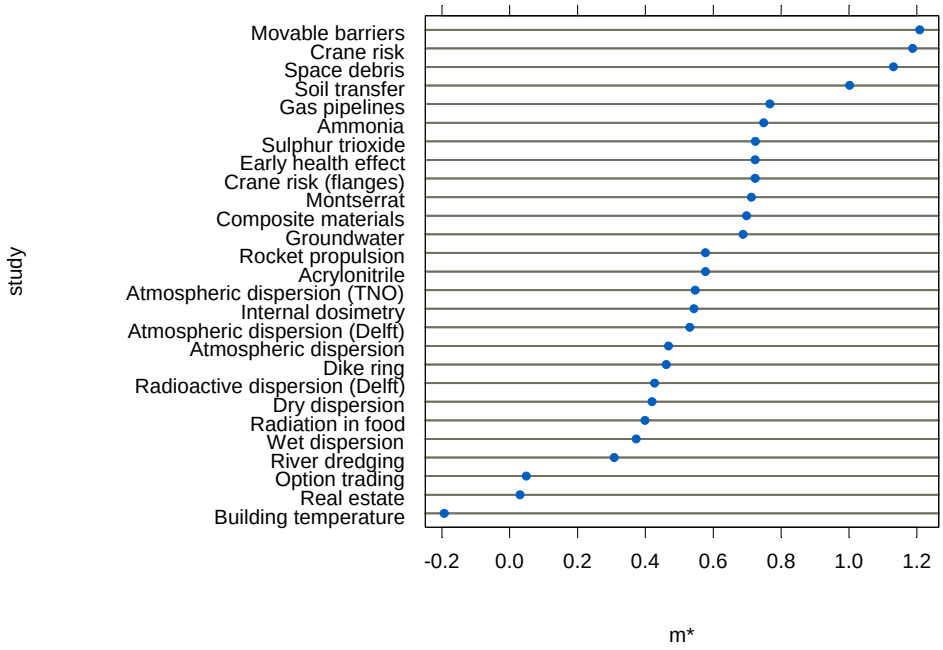

圖 3　不同專家小組在 *m 校準衡量下的總體表現**

由於我們希望進一步了解校準變異究竟是源自於專家之間的差異（例如源自於專業能力或知識上的不同、機率估計能力的落差，甚或在校準與明確度之間的權衡），或者是來自於問題之間的難易不同。因此，我們在圖 4 中呈現了不同研究中個別專家的 *m** 校準分數，圖中指出在同一個研究中，不同專家間的平均校準差異頗大，但是在不同的研究（專家小組）之間，專家校準分數的變異程度則似乎沒有太大的落差，基本上，並沒有出現某些小組中專家的表現很一致地達到完美校準，而其餘小組就出現參差不齊或變異很大的現象。另外，我們亦發現，雖然以理想值 0 作爲分界時，多數專家的判斷呈現過度自信，但是對分屬不同領域的每一研究來說，往往都有少數的專家擁有接近完美校準的表現。

我們也在圖 5 呈現了不同研究中，個別問題的 *m** 校準分數。圖中指出不同問題的校準分數差異頗大，這表示，校準的差異不只來自於不同研究領域的特性（例如有些領域可能必須依賴大量的實驗或者實證數據，而另一些領域則著重在理論與計算模型），某些問題似乎對專家而言較爲容易，也因此較能夠促成完美校準，而部份問題甚至造成自信不足的現象。另外，透過仔細比較圖 4 與圖 5，我們也發現不同問題之間校準分數的變異比不同專家之間的變異大。當然，上述透過圖形的觀察未必十分可靠，因爲這些研究中專家小組的成員並非很多（大多是由小於 10 人的專家組成），而且大多數的區間估計只能夠找到唯一的實現值，因此校準分數亦受到此實現值之內在隨機性的影響，因此以下我們將透過較嚴謹的統計分析來驗證上述的觀察。

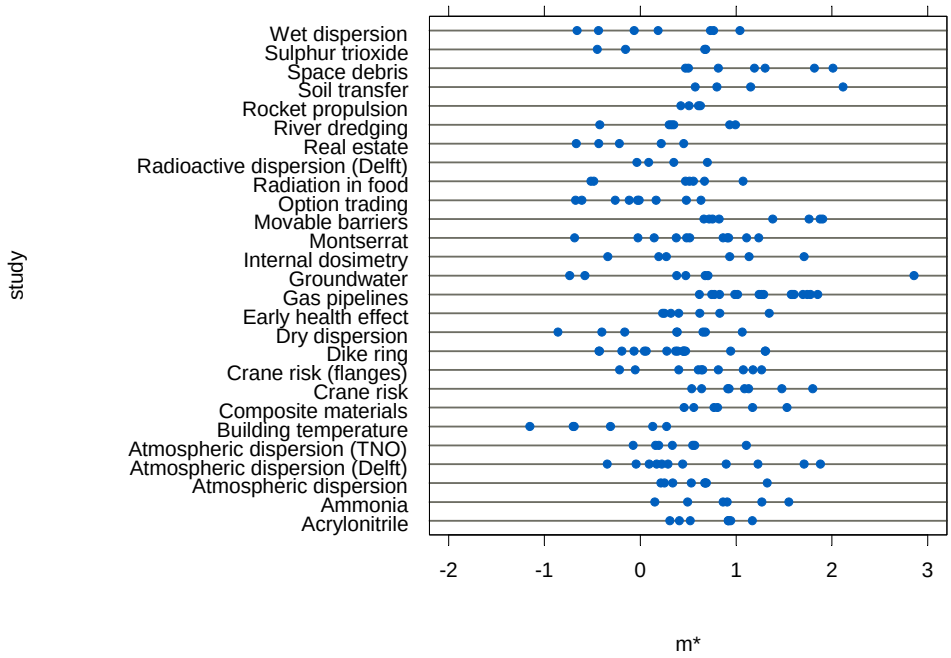

圖 4　不同專家的 *m 校準衡量**

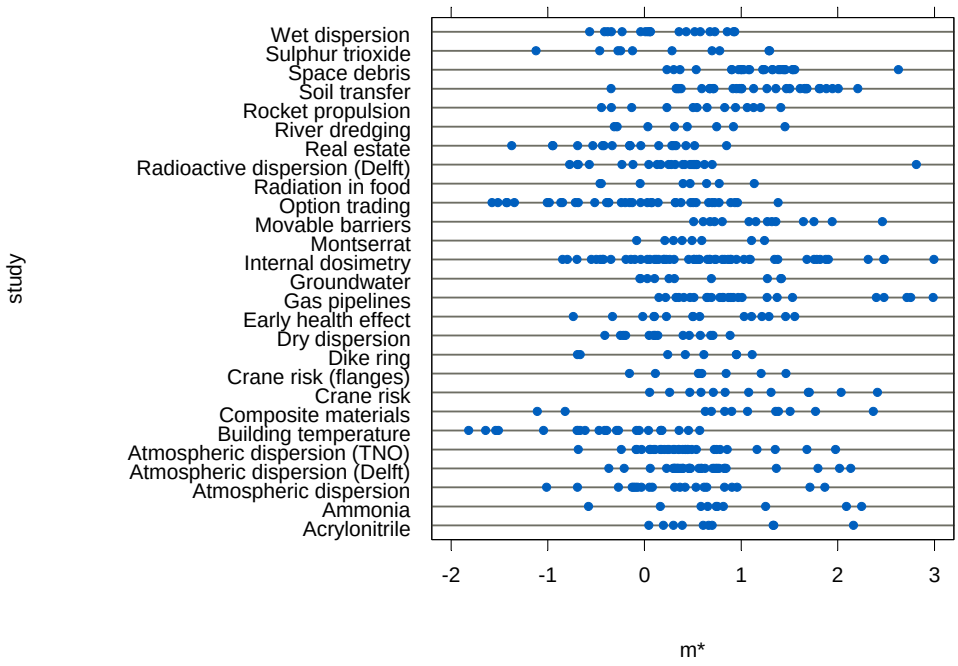

圖 5　不同問題的 *m 校準衡量**

4.3 線性混合模型統計分析

Lin and Bier（2008）的研究使用二元（binary）的反應變數來代表觀察到的實現值是否落在相對應的專家機率區間中，所以他們必須使用統計原理與參數估計均較爲複雜的邏輯常態廣義線性混合模型（logistic-normal generalized linear mixed model）。現在，由於我們已經推導出一個連續型的校準衡量 m^* (= log(m) = log(AD/EAD))，因此我們可以改用發展成熟的線性混合模型來進行統計分析。因此，當我們考慮第 i 個研究中的第 j 個專家，在研究過程中針對第 k 個問題提出區間估計，而此問題擁有 l 個實現值，則我們的線性混合模型（LMM）可以表示爲：

$$\eta_{l(kj(i))} = m^*_{l(kj(i))} = \left(study\ \text{effect}\right)_i + \left(expert\ \text{effect}\right)_{j(i)} + \left(question\ \text{effect}\right)_{k(i)} + residual_{l(kj(i))} \tag{9}$$

和之前 Lin and Bier（2008）使用的 GLMM 相似，由於在每個不同（領域）的研究中所使用的專家與問題並不相同，因此在統計實驗設計與分析上，這樣子的資料型態是屬於巢式設計（nested desgin），其中，「專家」與「問題」乃套疊於「研究」之下。因此，式子（9）中的 $j(i)$ 是用來表示 j 效應乃是套疊於 i 效應之中，而 $m^*_{l(kj(i))}$ 則服從於常態分配。爲了和之前的 GLMM 模型進行比較，在這個 LMM 模型中，我們仍將不同的研究對專家校準的影響設定爲固定效應（fixed effect），因此可以用來提供每個研究校準高低的估計。不過，專家、問題和殘差三者的影響都被假設爲隨機效應（random effect），且各自的分配爲 N(0, $I\sigma^2_{expert}$), N(0, $I\sigma^2_{question}$), 以及 N（0, $I\sigma^2_{residual}$）。

與 GLMM 模型對照下，我們發現 LMM 模型不再存有不同「實現值」之間的隨機效果，而是以「殘差」效果代替。然則，從資料與模型的結構看來，我們知道「實現值」的效果和「殘差」其實是不可分的（confounded），而他們在詮釋上也具有相同的義意。我們依據（9），透過 SAS Mixed 程序，進行模型分析與參數估計，所得到的結果如表 2 所示，隨機效果的參數估計值分別爲 σ^2_{expert}=0.236， $\sigma^2_{question}$=0.340，以及 $\sigma^2_{residual}$=1.272，統計分析並指出隨機效果均顯著不爲 0。另外在固定效果的檢定中，我們亦可以清楚的發現在研究之間校準度的差異無疑是十分顯著的（$p < 0.0001$）。

對於模型的進一步詮釋，我們著重在比較這幾個變異量成分（variance component）的大小，我們雖然由圖 4 及 5 中發現不同專家與不同問題的校準值之間存在相當大的變異，但統計分析的結果指出大部分的變異會被不同實現值之間的隨機效果所解釋，問題與專家的隨機效果則相對較小，且專家之間的變異又約只有問題之間變異的 2/3，符合我們在初步圖形分析中的觀察。然而，因爲在我們的資料庫中，只有部分問題擁有多個實現值，因此在統計分析中參數估計的過程裡，這些問題將對不同實現值之間隨機效應之大小有決定性的影響。因此這部分的結論，

表2 LMM變異量成分之參數估計

隨機效果	估計值
expert(study)	0.236
question(study)	0.340
Residual	1.272

仍需後續的研究與更多的資料來確認其穩健性。

最後，我們試著將 LMM 與 GLMM 兩模型所得的數據放在一起比較（如表 3），由於這兩個不同模型的分析乃是基於不同尺度的衡量，因此直接比較隨機效果的大小並沒有太大意義，但是我們仍可以比較不同隨機效果在同一模型中的相對大小來理解校準差異的主要來源。我們發現若和其他的變異來源比起來，LMM 模型中專家的相對影響 — 亦即專家與專家之間的變異性 — 遠比 GLMM 模型中專家的相對效應小很多。而且我們發現在新的校準衡量下，即使納入了實現值之隨機效應，問題之間的變異性仍大於專家之間的變異性，這亦是兩個不同的模型之間的顯著不同。我們發現不論是使用二元回應變數的 GLMM 模型或者連續型校準衡量的 LMM 模型，問題或者專家之間的變異多可以被實現值的隨機變異所解釋，但在 LMM 中，隨機變異所能夠解釋的部分比例更高。

5. 結論與建議

由於二元校準衡量的不穩定性除了可能造成模型參數估計時的收斂問題，也造成我們對模型與資料的配適程度的質疑，爲了徹底解決這些問題，以及確認過去研究中所得之結論在不同的校準衡量下是否依然有效，我們先以 Soll and Klayman（2004）所定義的校準衡量 M 來探索

表3 各模型之間變異數參數估計之比較

模型	隨機效果	σ^2_{expert}	$\sigma^2_{question}$	$\sigma^2_{realization}$	$\sigma^2_{residual}$
Lin 和 Bier（GLMM）		1.281	1.360	—	—
簡化模型*		(0.178)	(0.146)		
Lin 和 Bier（GLMM）		1.475	0.633	1.131	—
完整模型*		(0.201)	(0.176)	(0.183)	
LMM 模型		0.236	0.340	—	1.272
		(0.032)	(0.033)		(0.029)

* Lin and Bier (2008)

不同研究中過度自信的現象。同時，爲了更進一步探索校準變異的來源，我們以 *M* 校準衡量爲基礎，發展出可以評估單一區間估計的新校準衡量 *m**，由於 *m** 是一個連續（而非二元）的測度，所以我們改用線性混合模型（LMM）進行分析。

研究中發現，若將 Lin and Bier (2008) 的二元校準衡量視爲校標，則不論是 Soll and Klayman (2004) 所定義的 *M* 校準衡量，或者我們在本研究中定義的 *m** 校準衡量，都具有一定的效度。所以當發現這些衡量不約而同地指出不同研究間的平均校準差異頗大，而且多數的研究中存在過度自信時，我們並未感到訝異。

然而，除了顯著的差異存在於不同的研究之間，本研究亦發現專家與問題的隨機效應亦是顯著的，同時，由於部分問題存在多個實現值，所以允許我們進一步評估區間估計的校準變異究竟主要來自於專家或者問題本身，還是不同實現值之間的隨機效應。本研究的最重要發現來自於線性混合模型（LMM）在新校準衡量 *m**下的分析結果，我們發現，若使用更穩健的連續型校準衡量 *m**，專家之間的變異約是問題之間變異的 2/3，而隨機的變異（即不同實現值所帶來的影響）則幾乎達到問題之間變異的四倍，即使我們把少數較極端的觀測值移除以進行相關的敏感度分析，分析的結果仍未出現太大改變。亦即即使我們將不同實現值的隨機影響納入考慮，在新的校準衡量下，問題之間的變異性仍然大於專家之間的變異性，這是我們的結果和過去研究之間的顯著不同。

雖然專家之間仍存有顯著變異性，顯示部份專家的校準表現相對較突出，因此支持使用加權平均來整合專家意見的作法，而這些方法（例如 Cooke, 1991）則會給予在種子問題的評估中表現優良的專家較高的權重。然而，由於本研究同時指出專家之間的變異相對小於其他如問題之間或者實現值之間的隨機變異，因此，使用種子問題來篩選較佳專家的作法的實際果效可能相對有限。是以在專家判斷的實務應用上，使用較佳的擷取程序來減少過度自信偏誤的發生與發展更穩健的專家整合方法來統整不一致的區間估計仍是不可或缺。

同時，由於問題之間的變異遠大於專家之間的變異，過去認爲問題間難度不同所形成的校準差異其實可能可以被實現值中的隨機效果所解釋，因而認爲專家在一個特定領域研究中「不同問題」的校準表現均會相當一致的想法就不再成立了。這個情況將可能不利於使用種子問題來篩選較佳專家的作法，因爲如果已知這些種子問題間的校準存在很大落差，那我們如何保證在種子問題的校準表現能夠有效反應在真實問題之上？同時，針對某特定研究中的問題，假如每個專家在回答這些問題的表現皆不理想的話，則加權平均專家意見的效果亦將十分有限，此時，使用較複雜的貝氏模型來加寬專家的區間估計或許是一可行替代方案。

本研究雖然不是著眼於擷取或整合專家機率判斷方法上的更新與突破，但是藉由專家判斷的實證資料來驗證其過度自信的程度與來源，並且從管理的角度檢視它們對整合專家意見時的影響，從決策分析方法論的觀點看來，本研究除了點出使用主觀機率判斷的實務操作時不可輕

忽的過度自信現象之外，也對過去常用整合專家機率判斷的不同方法的適用性提出佐證與修正之建議。另外，許多學者專家在看待決策問題時，往往假設其中的機率估計爲已知或者可以輕易取得，而一些管理科學教科書在決策分析部分見樹不見林的探討更強化了如是觀點（Chelst，1998），這個邏輯的謬誤，也可藉此實證研究進一步釐清。

本研究仍存在一些有待後續研究來加以突破的研究限制，首先，在我們的資料中，僅有部分的種子問題擁有多個實現值，這除了可能造成模型參數估計上些微的偏誤之外，也可能影響我們統計推論與一般化研究結論的強度。在未來的研究中，使用更多擁有多個實現值的問題（甚或在應用專家意見的實務研究中刻意加入這類問題）將可以幫助我們更清楚釐清校準變異的來源。另外，在我們的研究中，由於無法精確掌握不同研究的細部特性、專家的人口變數與其他特徵或者問題描述的清晰程度與難易度等，使得我們無法更深入地去解釋造成校準高低的成因，未來的研究可由這個面向探討一些管理上可操控的特性與校準之間的關係，進而提供擷取與使用專家機率區間估計的操作原則。

參考文獻

Baginski, S. P., Conrad, E. J., and Hassell, J. M., “The Effects of Management Forecast Precision on Equity Pricing and on the Assessment of Earnings Uncertainty,” *The Accounting Review*, Vol. 68, No. 4, 1993, pp. 913-927.

Budescu, D. V., Erev, I., and Wallsten, T. S., “On the Importance of Random Error in the Study of Probability Judgment. Part I: New Theoretical Developments,” *Journal of Behavioral Decision Making*, Vol. 10, No. 3, 1997a, pp. 157-171.

Budescu, D. V., Wallsten, T. S., and Au, W. T., “On the Importance of Random Error in the Study of Probability Judgment. Part II: Applying the Stochastic Judgment Model to Detect Systematic Trends,” *Journal of Behavioral Decision Making*, Vol. 10, No. 3, 1997b, pp. 173-188.

Chatfield, C., “Prediction Intervals for Time-series Forecasting,” In J. S. Amstrong (Eds.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell, MA: Kluwer Academic Publishers, 2001, pp. 475-494.

Chelst, K., “Can’t See the Forest Because of the Decision Trees: A Critique of Decision Analysis in Survey Texts,” *Interfaces*, Vol. 28, No. 2, 1998, pp. 80-98.

Cooke, R. M., *Experts in Uncertainty: Opinion and Subjective Probability in Science*, New York: Oxford University Press, 1991.

de Finetti, B., *Theory of Probability,* Volume I, New York: John Wiley & Sons, Inc., 1974.

Hora, S. C., "Probability Judgments for Continuous Quantities: Linear Combinations and Calibration," *Management Science*, Vol. 50, No. 5, 2004, pp. 597-604.

Johnson, W. B., "The Impact of Confidence Interval Information on Probability Judgments," *Accounting, Organizations and Society*, Vol. 7, No. 4, 1982, pp. 349-367.

Juslin, P., Wennerholm, P., and Olsson, H., "Format Dependence in Subjective Probability Calibration," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 25, No. 4, 1999, pp. 1038-1052.

Juslin, P., Winman, A., and Olsson, H., "Naïve Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect," *Psychological Review*, Vol. 107, No. 2, 2000, pp. 384-396.

Kahneman, D. and Riepe, M., "Aspects of Investor Psychology," *Journal of Portfolio Management*, Vol. 24, No. 4, 1998, pp. 52-65.

Klayman, J., Soll, J., González-Vallejo, C., and Barlas, S., "Overconfidence: It Depends on How, What and Whom You Ask," *Organizational Behavior and Human Decision Processes*, Vol. 79, No. 3, 1999, pp. 216-247.

Lichtenstein, S. and Fischhoff, B., "Do Those Who Know More Also Know More about How Much Rhey Know? The Calibration of Probability Judgments," *Organizational Behavior and Human Performance*, Vol. 20, No. 2, 1977, pp. 159-183.

Lichtenstein, S., Fischhoff, B., and Philips, L. D., "Calibration of Probabilities: The State of Art to 1980," In D. Kahneman, P. Slovic, and A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and Biases, Cambridge, England: Cambridge University Press, 1982, pp. 306-334.

Lin, S. W. and Bier, V., "A Study of Expert Overconfidence," *Reliability Engineering and System Safety*, 2008, in press.

Morgan, M. G. and Henrion, M., *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge, England: Cambridge University Press, 1990.

Plous, S., *The Psychology of Judgment and Decision Making*, New York: McGraw-Hill, 1993.

Russo, J. E. and Schoemaker, P. J. H., "Managing Overconfidence," *Sloan Management Review*, Vol. 33, No. 2, 1992, pp. 7-17.

Soll, J. B. and Klayman, J., "Overconfidence in Interval Estimates," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 30, No. 2, 2004, pp. 299-314.

Shlyakhter, A. I., Kammen, D. M., Broido, C. L., and Wilson, R., "Quantifying the Credibility of Energy Projections from Trends in Past Data – The United States Energy Sector," *Energy Policy*,

Vol. 22, No. 2, 1994, pp. 119-130.

Walley, P., *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall, 1991.

Wiper, M. P., French, S., and Cooke, R., "Hypothesis-based Calibration Scores," *The Statistician*, Vol. 43, No. 2, 1994, pp. 231-236.