國立臺灣大學文學院圖書資訊學系暨研究所

博士論文

Department of Library and Information Science

College of Liberal Arts

National Taiwan University

Doctoral Dissertation

同儕審查的評審標準、信度與公平性研究：以台灣出版之社會暨人文科學期刊為例

Criteria, Reliability, and Fairness in Peer Review: A Study of Taiwanese Social Science and Humanities Journals

嚴竹蓮

Chu-Lien Yen

指導教授：黃慕萱 博士

Advisor: Mu-Hsuan Huang, Ph.D.

中華民國 105 年 7 月

July 2016

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

國立臺灣大學博士學位論文
口試委員會審定書

同儕審查的評審標準、信度與公平性研究：以台灣出版之社會暨人文科學期刊為例
Criteria, Reliability, and Fairness in Peer Review: A Study of Taiwanese Social Science and Humanities Journals

本論文係嚴竹蓮君（學號 D96126005）在國立臺灣大學圖書資訊學研究所完成之博士學位論文，於民國一〇五年七月廿六日承下列考試委員審查通過及口試及格，特此證明

口試委員：

	（指導教授簽名）

系主任、所長 （簽名）

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

誌謝與感言

謝謝黃慕萱教授的指導及諸位論文口試委員的卓見,讓我的研究成果更堅實；也感謝致力於同儕審查研究的學界先進，因為有您們的肩膀，讓我能看得更遠！

目前我國同儕審查研究尚在起步階段，謹借拙著拋磚引玉，期待我國政府與學術界都能：

慎視同儕審查為知識研究的主體；

積極參與同儕審查的國際合作；

發展同儕審查的監督與檢驗機制；

進而營造一個開放的同儕審查程序、促進學術資源分配效率，以及健全學術研究環境。

doi:10.6342/NTU201602255

Acknowledgements

I wish to express my sincere gratitude to my academic advisor, Professor Mu-Hsuan Huang, and all the other members of my dissertation committee, for their generous advice and guidance, which have made my research more relevant and substantial. I would also like to thank members of academia worldwide for their dedicated study and research on peer review. It is by standing on the shoulders of giants that I have been able to see further.

Research on peer review is relatively new to Taiwan. This study marks an effort to bring this topic into the spotlight and encourage our government and academic community to:

consider peer review as a branch of academic knowledge,
participate in related international projects,
and establish a monitoring and evaluation mechanism,

so as to develop an open and transparent procedure for peer review, ensure efficient allocation of academic resources, and create a healthy environment for academic research.

doi:10.6342/NTU201602255

摘　　要

同儕審查是學術界進行科學探索時所採用的一項自律機制，幾乎已制度化地納入學術組織的運作之中，並普遍獲得學界人士的支持。基本上同儕審查的正當性是基於學術社群成員之間的信賴與誠信，在各項學術活動中以不同的作業模式分配有限資源，包括學術文獻出版、研究計畫獎助、大學教職聘用與升遷，以及學術成就獎勵等。但是同儕審查的運作方式迄今未臻完善，除了出現效用、效率，以及信度等問題外，許多研究亦已證實存在多種評審偏見，包括機構偏見、交情偏見、年齡偏見，以及保守偏見等。

本研究之研究目的有三，其一為回顧同儕審查機制之起源與發展，並分析同儕審查做為知識研究主體的歷程，綜整結論如下：同儕審查評審機制逐漸開放與透明、同儕審查的國際交流持續進行、同儕審查評審品質與效能提升的作法多元，以及同儕審查評審作業的未來發展，包括建立持續監督、檢驗與改進的同儕審查機制及同儕審查與書目計量的競合關係。其二是探討同儕審查之評審標準、信度與公平性的研究現況，綜整結論包括：評審標準的研究依然受到重視、評審信度過低的因果研究有待加強、公平性研究的方法論受到質疑，以及同儕審查的效用仍待證實。

另外本研究亦以 3 種我國出版之社會暨人文科學期刊之評審報告進行實徵研究，探討評審者在審查過程中使用的實際標準，並檢測評審者彼此之間的信度，以及討論評審的公平性與課責性議題，研究發現：評審報告的內容分析呈現評審者實際使用之核心標準、出版建議分析可細部解讀拒絕稿件或接受稿件之主要理由、稿件研究類型及稿件領域分析呈現非量化研究及人文科學研究的評審特性、公平性議題需審慎推論因果、評審評語與出版建議的一致性應進行系統討論、評審者行為理論之研究有其必要性，以及我國人文科學研究有其特殊的撰稿方式。

綜合而言，學術領域越來越專精且複雜，研究人口也愈來愈多，在學術資源未見大幅成長的背景下，競爭將日趨激烈，同儕審查研究的益受重視將不言可喻。

doi:10.6342/NTU201602255

證諸許多著名同儕審查期刊的稿件接受率只有個位數字，許多國家的政府獎助機構的獲獎率也有逐年降低的趨勢，某些頗具聲望的獎學金甚至僅有 1/200 的機率；另外各國高教經費緊縮及學生人數減少，也助長了大學教職之路的競爭。因此未來同儕審查的作業勢必面對外界更多的質疑與挑戰，而建立一個持續性監督與檢驗的機制是學術界的努力方向，進而營造一個開放的同儕審查程序、促進學術資源分配效率，以及健全學術研究環境。

關鍵詞：同儕審查、書目計量

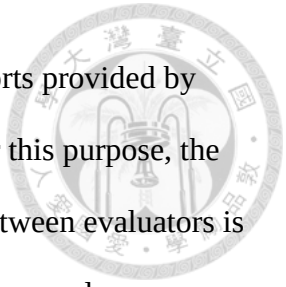
doi:10.6342/NTU201602255

Abstract

Peer review is a self-regulation mechanism for scientific inquiry. Institutionalized and incorporated into the structure and operation of science, it has received considerable support in the academic setting. The legitimacy of peer review is based on trust and integrity. In various ways, it allocates scarce resources such as journal space, research funding, faculty recruitment, career advancement, and rewards for academic achievements. But there are growing indications of unresolved deficiencies in the operation of peer review, leading to negative assessments as to whether it is effective, efficient, or reliable. Many studies have found links between potential sources of bias and judgments in peer review, such as institutional prejudice, cronyism, ageism, and conservatism.

This study aims to achieve three objectives. First, it explores the origins of peer review and traces the process by which it has become a subject of academic research. This examination shows that peer review has gradually become more open and transparent, has inspired ongoing international exchange, and has embraced diverse approaches for higher-quality evaluation. There is also increased anticipation for mechanisms to be established for continual supervision, scrutiny, and improvement, as well as for a competitive–cooperative relationship to develop between peer review and bibliometrics. Second, this study seeks deeper insights into contemporary research on different assessment criteria and on the reliability and fairness of peer review. While review criteria continue to be a major focus of research, there is also a need for greater investigation of the reasons for low inter-evaluator reliability, doubt has been cast on the methodology of studies on the fairness of peer review, and the effectiveness of peer review remains to be demonstrated.

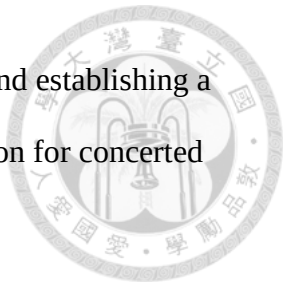
doi:10.6342/NTU201602255



Third, this paper presents an empirical study of peer review reports provided by three social science and humanities journals published in Taiwan. For this purpose, the actual criteria employed are examined and the degree of reliability between evaluators is scrutinized; this is accompanied by a discussion on the issues of fairness and accountability in peer review. The findings include: the core criteria employed by evaluators, as revealed by content analysis of review reports; the main reasons for manuscript acceptance or rejection, as revealed by detailed interpretation of publishing recommendations; and the special attributes of peer review in non-quantitative and humanities research, as revealed by analysis of manuscripts in different research categories and specialized domains. These findings also highlight the need for caution in inferring cause and effect with regard to issues of fairness; for systematic debate with regard to consistency between evaluators' review comments and their publishing recommendations; for research into the behavior of evaluators; and for consideration of the specific characteristics of manuscript preparation in humanities fields in Taiwan.

Overall, academic disciplines are becoming increasingly specialized and complex, and the research population is growing ever larger, yet there has been no great increase in academic resources. In these circumstances, competition will grow increasingly intense, and thus it seems likely that research into peer review will attract ever greater attention. This growing competition is evidenced by the facts that acceptance rates for manuscripts submitted to many leading peer-reviewed journals are in single figures, and that rates of funding allocation by government funding organizations in various countries are also trending downward year by year, falling to a ratio of 1:200 at some renowned institutions. Budgetary reductions and declining numbers of students in higher education have also heightened competition among faculty. Therefore, peer review practice will inevitably face increasing questions and challenges from the

doi:10.6342/NTU201602255



outside world. Creating an environment for transparent peer review and establishing a mechanism for continual supervision and scrutiny should be a direction for concerted effort within academia.

Keywords: peer review, bibliometrics

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

目　　次

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

圖目次

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

表目次

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

第一章　緒論

第一節　問題陳述

「同儕審查」（peer review）是當代學術領域中相對具有發展性的研究主題，它原本用以輔助科學探索，容許科學家運用它來評判某一特定科學研究的品質，以產生具有信度與效度的知識。之後同儕審查的應用範圍逐漸擴大，成為學術界分配有限資源的主要機制之一，在期刊文獻出版、獎助資源分配，以及大學教職的聘用與升遷上都扮演著關鍵角色（Bornmann, 2011b; Frodeman, Holbrook, & Mitcham, 2012; Kronick, 1990; Langfeldt & Kyvik, 2011）。近半個世紀以來，同儕審查的效用（effectiveness）、效率（efficiency）、信度（reliability），以及公平性（fairness）等受到多方質疑，有關同儕審查的研究也因此日益受到重視，研究的廣度與深度亦逐漸增加，而成為一個獨立且益形重要的學門。

同儕審查相應的英文同義詞除了 peer review 之外，還包括 peer advice、peer evaluation、peer judgement、peer censorship、merit review，以及 refereeing 等（Chubin & Hackett, 1990）。中文也有不同的說法，例如同儕審查、同儕互評、同儕評論、同儕評鑑、同儕評估、同行評議、同行評審，以及同行評閱等，本研究將採用「同儕審查」做為前述中、英文含義的通用詞彙。

基本上同儕審查是一個通稱，在不同的機構有不同的作法，甚至同一機構的不同單位亦可能有所差異（General Accounting Office [GAO], 1999; Organisation for Economic Co-operation and Development [OECD], 2011a; Wood & Wessely, 2003）。Chubin 與 Hackett（1990, pp. 1-2）將之定義為：「同儕審查是一套科學研究的品質評價機制，科學界利用同儕審查來確認研究程序的正確性及推論的合理性，並根據評審結果分配有限資源，例如期刊版面、獎助名額，以及學術聲望或特殊榮譽等。」近年來同儕審查已廣泛應用於學術研究的各個層面，本研究綜整學者意見

doi:10.6342/NTU201602255

將同儕審查分為三大類：（一）出版品同儕審查：包括文獻類出版品（如期刊文獻及專書等）及非文獻類出版品（如紀錄片、資料庫、網站，以及電腦軟體等）的同儕審查；（二）獎助同儕審查：包括獎助計畫及獎學金的同儕審查；（三）成就同儕審查：評鑑個人、團隊、部門或機構的學術表現，並依據評審結果分配資源或獎勵，包括大學教職聘用／升遷同儕審查、學術榮譽或獎項同儕審查，以及教育暨研究機構評鑑等（Harley & Acord, 2011; Parliamentary Office of Science and Technology [POST], 2002; Research Information Network [RIN], 2010）。

許多學者認為同儕審查可上溯至 1665 年英國倫敦皇家學會（Royal Society of London）創辦的《哲學學報》（*Philosophical Transactions*），當時每期學報在出刊前，均須經由學會會員審查內容，而被視為期刊同儕審查的濫觴（Burnham, 1990; Kronick, 1990; Lock, 1985; Rennie, 2003; Spier, 2002a）。之後隨著科學研究的機構化及學術期刊的市場化，期刊稿件的同儕審查不僅是科學知識傳播的守門人，也成為科學研究品質的最後仲裁者（Biagioli, 2002; Burnham, 1990; Kronik, 1990; Lock, 1985; Marsh, Jayasinghe, & Bond, 2008; Rennie, 2003; Spier, 2002a; Weller, 2002; Zuckerman & Merton, 1971）。正如國際醫學期刊編輯委員會（International Committee of Medical Journal Editors, ICMJE）訂定之醫學期刊編輯與出版建議規範指出：「（期刊）同儕審查以其公平性、獨立性，以及批判性的特質，成為科學研究過程的重要延伸環節」（International Committee of Medical Journal Editors [ICMJE], 2013, p. 5）。

獎助同儕審查的發展與政府職能擴張的關係密切，當西方各國政府開始擔負起支持科學研究的責任，同儕審查就逐漸成為政府獎助經費的分配工具；尤其在第二次世界大戰結束後，英美等國積極支持科學研究，同儕審查即成為政府獎助研究機構的主要經費分配機制，以提升政府科研投資的效率與效能為目標（Burnham, Sauer, & Gibbs, 1987; Frodeman et al., 2012）。至於大學教職聘用／升遷同儕審查則是在全球高等教育普及化及大學教職聘用制度化的環境下，逐漸成為

doi:10.6342/NTU201602255

學者進入校園任教的仲裁者（Snodgrass, 2006; Weiser, 2012）。20 世紀後期，英美等國政府強調科學管理，積極主導國家科學發展方向，同儕審查遂進一步成為科技政策制定及教研機構評鑑的重要工具（Frodeman & Briggle, 2012; Guston, 2003; Whitley & Gläser, 2007）。根據經濟合作暨發展組織（OECD, 2011a）的同儕審查專刊稱，在知識經濟的社會，科學研究在國家創新過程中的角色愈來愈重要，更是政府在進行政治、經濟，以及社會決策時的重要參考，因此使得同儕審查的評審標準與審查過程也更加複雜。

300 多年來，同儕審查受到學術界的重視，日益擴大應用範圍，但是批評聲浪亦眾，許多文獻質疑同儕審查的真實效用、批評其過程耗錢費時，並且發現評審不公的現象，包括機構（institutional）偏見、交情（cronyism）偏見、年齡偏見、性別偏見、國籍偏見、非英語母語偏見、保守（conservative）偏見、學派偏見、學術產出偏見（productivism, 強調研究文獻出版），以及審查先後順序偏見等（Abdoul, Perrey, Amiel, et al., 2012; Bornmann, 2011b; Fang, 2011; Lee, Sugimoto, Zhang, & Cronin, 2013; Rennie, 2003; Sandström & Hällsten, 2007; Wenneras & Wold, 1997; Wood & Wessely, 2003）。儘管如此，大部分學者依然認為有必要維持這個機制，因為其他替代方案的爭議性更大；有些學者甚至將同儕審查類比為自由世界的民主制度，是不好機制中的最佳選項（Harley, Acord, Earl-Novell, Lawrence, & King, 2010; Ismail, Farrands, & Wooding, 2009; Kostoff, 2004; Rennie, 1986; Sieber, 2006）。

雖然學術界早已認知到同儕審查作業的不完美，但是直到 1980 年代前後，同儕審查的理性檢驗才逐漸受到重視，1989 年期刊同儕審查開始定期舉辦研討會，也帶動了一股研究風潮（Chubin & Hackett, 1990; Rennie, 2003; Weller, 2002）。目前同儕審查的實徵研究大多以期刊稿件為主，獎助計畫次之（Bornmann, 2011b; Demicheli & Di Pietrantonj, 2007; Jefferson, Rudin, Brodney-Folse, & Davidoff, 2007; Weller, 2002; Wood & Wessely, 2003），大學教職聘用／升遷同儕審查的研究不多

doi:10.6342/NTU201602255

（Miller, 1978; Weiser, 2012）。Bornmann（2011b）回顧近十餘年的期刊及獎助同儕審查實徵研究，其中大多數為信度及公平性議題，至於預期效度（predictive validity）的研究較少。不過綜合來看，許多文獻的方法論不夠嚴謹、因果推論不足，而且各篇研究結果的異質性甚大難以通則化。因此一些學者建議進行實驗性研究以強化因果分析，並利用後設分析法解決通則化的問題（Bornmann, 2011a; Bornmann, Mutz, & Daniel, 2007; Bornmann, Nast, & Daniel, 2008; Dalton, 1995; Wood & Wessely, 2003），然而 Marsh、Jayasinghe 與 Bond（2011）卻認為，除了二手文獻的後設分析外，大規模一手評審資料的研究亦有其必要性。

無論是何種類型的同儕審查，大多有個非常隱密且幾乎不透明的程序，因而被稱之為「黑盒子」，許多學者認為同儕審查研究的困境主要來自於此。Chubin 與 Hackett（1990）指出，只有少數文獻利用期刊稿件評審者的審查報告進行研究，同樣的情況也出現在獎助同儕審查，大多數的學者無法直接由獎助機構取得申請者、評審者或評審過程相關資訊。直到 20 世紀末期，許多國家的公共獎助機構在政府資訊開放及政策目標管考的壓力下，為了提升同儕審查的作業品質與效能，比較願意將第一手評審檔案提供學者進行研究，有些甚至允許學者訪談評審者或觀察評審委員會的議事與決策過程（Bornmann & Daniel, 2005; Lamont, 2009; Langfeldt, 2001; Marsh et al., 2008; Wood & Wessely, 2003），使得獎助同儕審查的評審作業逐步邁向開放與透明之途。但是截至今天，大學教職聘用／升遷同儕審查的評審過程，仍然是個密閉的黑盒子，研究資料取得最為不易。

有關同儕審查的評審標準研究開始較早，文獻篇數也較多（Chubin & Hackett, 1990）。理論上評審標準研究最直接的方法是利用評審者的審查報告，或是評審委員會的實況觀察或是會議紀錄進行分析（Abdoul, Perrey, Amiel, et al., 2012; Bornmann & Daniel, 2004, 2005），以瞭解操作時使用的實際標準（practical criteria or actual criteria）；但是無論是期刊或獎助同儕審查，此類研究均不多見。以期刊稿件的評審標準研究來看，大多係針對期刊編輯、評審者、稿件作者或一般學者

doi:10.6342/NTU201602255

進行的調查訪問，有學者將這類研究所得稱之為規範標準（normative criteria or theoretical criteria）（Chubin & Hackett, 1990; Weller, 2002），以與操作上的實際標準有所區隔。至於獎助計畫的評審標準研究，最初亦多是針對同儕審查的參與者進行意見調查，或者蒐集獎助機構訂定的評審標準加以分析，亦均屬於規範標準的性質。直到 20 世紀末期，許多國家的政府獎助機構陸續同意公開評審過程檔案，獎助同儕審查的研究邁向了實際標準的時代，有學者觀察獎助評審委員會的議事運作及決策過程（Lamont, 2009; Langfeldt, 2001, 2006; Luukkonen, 2012; Olbrecht & Bornmann, 2010）、有學者分析內外部評審者的審查報告及決審會議紀錄內容（Abdoul, Perrey, Amiel, et al., 2012），亦有學者探討不同評審階段的評審者所使用的評審標準，以及各階段評審結果的穩定性與相關性等（Bornmann & Daniel, 2005; Bornmann, Mutz, & Daniel, 2008; van Arensbergen, van der Weijden, & van den Besselaar, 2014a; van den Besselaar & Leydesdorff, 2009）。目前大部分獎助同儕審查的研究指出，無論是內部或外部評審者或評審委員會成員多依據獎助機構所訂定之規範標準進行審查，但是每位評審對於各項規範標準的重視程度有別，是否因而影響評審結果，仍有待進一步研究（Abdoul, Perrey, Amiel, et al., 2012; Bornmann & Daniel, 2005; Langfeldt, 2001）。

除了評審標準之外，期刊與獎助同儕審查的評審信度與公平性也是受到學者關切的兩項議題。所謂評審信度是指評審者彼此之間的一致性，至於評審的公平性則是強調評審標準應與學術研究的科學品質與價值相關，不應受到作者／申請者的個人特殊條件或評審者偏見等因素的影響。許多研究已經證實同儕審查的評審信度過低，推論可能起因於評審者的偏見（Bornmann & Daniel, 2010; Campanario, 1998a, 1998b; Chubin & Hackett, 1990; Cicchetti, 1991; Langfeldt, 2001; Opthof & Wilde, 2009），不過也有學者認為評審間的低信度，符合科學審查之常態，研究前沿的低度共識即可見一斑（Bailar, 1991; Chubin & Hackett, 2003; Cole, 1992; Harnad, 1985; Hodgson, 1997; Rennie, 2003; Stricker, 1991）。另有文獻指出同儕審查存在評

doi:10.6342/NTU201602255

審不公的情況，最常出現的是機構偏見、交情偏見，以及性別偏見（Bornmann, 2011a; Marsh & Bornmann, 2009; Wood & Wessely, 2003）。有些學者建議利用評審者的審查報告進行分析，或許有助於釐清評審者之間信度過低的問題，以及強化評審公平性的因果探討（Bornmann, Mutz, & Daniel, 2010; Fiske & Fogg, 1990; Marsh et al., 2011; Siegelman, 1991）。

相較於期刊與獎助同儕審查，大學教職聘用／升遷同儕審查的研究甚少，有關評審標準的文獻亦只有少數調查性研究，對於評審的信度與公平性則未見相關討論。但是在 20 世紀後期，歐美各國大學生人數逐漸下降及政府高等教育預算減少，許多學院開始刪減教職員人數，使得大學教職的競爭益形激烈（Becher & Trowler, 2001; Weinbach & Randolph, 1984），因此大學教職聘用／升遷同儕審查的公平性逐漸受到關切。此外近年來各國政府認知到高等教育品質及學術研究創新的關鍵在於卓越的科學家，更加重視學術人力資源管理，有關學者個人表現的評審標準，也受到廣泛討論。歐盟第 7 架構計畫（EU 7th Framework Programme）所支持的學術生涯評鑑研究計畫（Academic Careers Understood through Measurement and Norm, ACUMEN），已提出個人學術表現評審架構－ACUMEN Portfolio，讓大學教職聘用／升遷同儕審查的評審標準，有了更開放且多元的想法（ACUMEN Consortium, 2014; Tatum & Wouters, 2013; van der Most, 2014）。

同儕審查是學術品質評鑑的主要機制之一，全球先進國家大都採用此一方式分配有限資源，根據相關回顧文獻發現，大多數的研究來自於歐洲、美加及澳洲等西方國家，亞洲國家對此一主題的研究較少，即使是在某些以科技創新著稱，而且高度倚重同儕審查進行學術資源分配的國家如日本、韓國及新加坡亦然。我國目前有關同儕審查的研究甚少，大多為論述式文獻，有分析同儕審查在期刊編輯與出版過程中所扮演的角色（卯靜儒，2013；林娟娟，1997）、有針對學術電子期刊同儕審查系統之探析（邱炯友，2003），還有一些討論評審者的審查倫理（陸偉明，2009；黃毅志、曾世杰，2008）。總之我國學術界對於同儕審查的批評仍多

doi:10.6342/NTU201602255

流於私下抱怨（陸偉明，2009），而未將之當成知識研究的主體。反之西方學者經過半世紀的努力，已將同儕審查視為一門學術知識，為其建構科學研究的價值與方向。

綜合而言，學術領域越來越專精且複雜，研究人口也愈來愈多，在學術資源未見大幅成長的背景下，競爭將日趨激烈，同儕審查研究的益受重視將不言可喻。證諸許多著名同儕審查期刊的稿件接受率只有個位數字，許多國家的政府獎助機構的獲獎率也有逐年降低的趨勢（National Institutes of Health [NIH], 2013; National Science Foundation [NSF], 2011; Powell, 2010; Research Councils UK [RCUK], 2006），某些頗具聲望的獎學金甚至僅有 1/200 的機率（Lamont, 2009），低到幾已無評審之必要性；另外各國高教經費緊縮及學生人數減少，也助長了大學教職之路的競爭（Bloch, Graversen, & Pedersen, 2014; Harley & Acord, 2011; Laudel & Glaser, 2012）。因此未來同儕審查的作業勢必面對外界更多的質疑與挑戰。

本研究首先探討同儕審查機制之起源、發展與研究現況，並以 3 種我國出版之社會暨人文科學期刊為例進行實徵研究，以內容分析法分析同儕審查之評審報告，探索評審者在審查過程中使用的實際標準，並檢測評審者彼此之間的信度，以及討論評審的公平性與課責性（accountability）議題。

本研究的特色有三，其一、我國首篇整體回顧同儕審查之起源、研究現況與發展之論文；其二、首篇以我國社會暨人文科學期刊的同儕審查評審報告進行之實徵研究；其三、研究設計有三大創新，一者利用稿件研究類型（量化研究與非量化研究）探討我國社會暨人文科學期刊的同儕審查模式；二者利用評審者之評審標準進行信度分析；三者設計評審者的評語評分系統，用以探討評審的課責性。此外在論文最後將針對我國同儕審查機制與研究的整體發展方向，以及我國社會暨人文科學期刊同儕審查作業的未來研究提出建議。

doi:10.6342/NTU201602255

第二節　研究目的

本研究之研究目的如下：

一、探討同儕審查機制的起源、檢驗與發展，分析同儕審查做為知識研究主體的歷程。

二、探討同儕審查之評審標準、信度與公平性的研究現況，並針對目前期刊同儕審查品質提升的各種作法進行分析。

三、利用我國社會暨人文科學期刊之同儕審查的評審報告進行實徵研究，並與目前相關文獻進行比較，研究重點如下：

（一）探討我國社會暨人文科學期刊同儕審查之評審標準。

（二）探討我國社會暨人文科學期刊同儕審查之評審信度。

（三）探討我國社會暨人文科學期刊同儕審查之評審的公平性。

（四）探討我國社會暨人文科學期刊同儕審查之評審的課責性。

第三節　研究範圍與限制

本研究在文獻回顧部分僅限於中文或英文之圖書、期刊文獻、會議論文，以及機構報告等，至於期刊同儕審查之實徵研究，係以我國出版之社會暨人文科學期刊的同儕審查評審報告為分析對象，研究的範圍與限制，以及比較文獻之選擇說明如後。

一、研究範圍

同儕審查的研究主題甚多，本研究僅針對期刊稿件同儕審查的評審標準、信度及公平性進行討論，另因評審公平性的因果推論不易，而提出評審的課責性議題，以實質強化期刊同儕審查之評審品質，達到提升評審公平性之目的。至於期

doi:10.6342/NTU201602255

刊同儕審查的效率（時間與成本）、預期效用、評審者的偵錯能力，以及有關剽竊研究創意等誠信議題均不在研究範圍之內。

二、研究限制

本研究採便利抽樣（convenience sampling），以我國出版的社會暨人文科學之A級或 B 級期刊為對象，並僅就各期刊所提供之評審者的審查報告及出版建議進行討論，有關評審者及作者之個人背景、稿件的主題，以及編輯的出版決定等，均不在資料蒐集範圍之內。

本研究主要探討期刊稿件評審者所使用之實際評審標準，並比較評審者之間的信度，以及評審的公平性，交叉分析變項包括評審者出版建議、稿件研究類型（量化研究及非量化研究），以及稿件領域（社會科學及人文科學）。至於稿件所屬學門，因單一樣本數有限而不予討論。

另為保護資料來源，本研究亦不針對個別期刊的特色如期刊之學科領域或類型（綜合或專門期刊）、編輯委員之組成方式、評審報告格式設計（是否有選單式評分等），以及期刊的拒稿率等項目進行分析或檢驗。

三、文獻比較

有關期刊同儕審查實徵研究的結果將與目前相關文獻進行比較，以呈現我國期刊同儕審查之特色。但是因為目前期刊同儕審查文獻的研究結論未有系統性發現，而且異質性甚大，是故比較文獻之選擇以後設分析研究或論述式的回顧文獻為主，個別研究資料為輔。

doi:10.6342/NTU201602255

第四節　名詞解釋

一、大學教職聘用／升遷同儕審查（faculty appointments and promotions peer review）

大學或學院等高等教育機構為教職聘用或升遷的申請案所採行之同儕審查作業，目前各學校之同儕審查作業方式多不相同，甚至同一所大學內的不同系所也可能有別，此外各系所也會依據遴聘時之選才特定需求而調整評審標準與作法（Frodeman et al., 2012; Gross-Schaefer, Gala, Jaccard, & Vetter, 2015; Weiser, 2012）。

二、公平性（fairness）

同儕審查是學術社群的自我規範機制，公平性則是同儕審查機制合理性的基礎。同儕審查的公平性研究是為了提升審查作業的公平性並減少評審偏見，（Bornmann, 2011a），也就是所謂的興利除弊。本研究的公平性係指評審結果與學術研究產出（或表現）的科學品質與價值相關，若是受到非科學品質與價值因素的影響則為不具公平性，例如受評者的性別、聲望或所屬機構等個人特殊條件，或是由評審者的個人偏見等。

三、同儕審查（peer review）

同儕審查是一個通稱，在不同的機構有不同的評審標準與作法。Chubin 與 Hackett（1990, pp. 1-2）的定義強調效用與應用：「同儕審查是一套科學研究的品質評價機制，科學界利用同儕審查來確認研究程序的正確性及推論的合理性，並根據評審結果分配有限資源，例如期刊版面、獎助名額、學術聲望與特殊榮譽等」（Chubin & Hackett, 1990, pp. 1-2）。而就操作層面來看，「Michael Gibbons 與 Luke Georghiou 認為同儕審查是某一特定領域的專家，對於相同或相近領域之其他科學

doi:10.6342/NTU201602255

家的研究作品，進行科學特定面向（如研究品質）的專業價值判斷。其前提係基於評審專家必須對領域的認知發展、研究方向，以及研究社群有充足的知識」（OECD, 2011a, p. 1）。

四、信度（reliability）

Cicchetti（1991）定義同儕審查的評審信度（inter-reviewer reliability, IRR）為：「對於同一科學文獻之兩份或兩份以上的獨立評審報告的一致程度（p. 120）。」許多學者利用 intraclass correlation coefficient（ICC）來測量評審信度，ICC 是一種變異數分解法，其值介於正 1 與負 1 之間，其缺點為高信度若加上評審者之間的低變異性，信度可能是基於偶然率。因此有些同儕審查的文獻利用 Kappa 係數來測量評審者之間的信度。所謂 Kappa 係數是有關兩位或兩位以上評審者之間的一致性，如果評審者的意見完全一致時則 K=1，如果 K 值接近 0，評審者之間的共識度不會高於偶然率。

五、評審標準（criteria）

同儕審查的評審者（期刊編輯、評審者或評審委員）在審查期刊稿件、獎助計畫，以及大學教職聘用／升遷案件時所採用的標準。一般分為規範標準（normative criteria or theoretical criteria）及實際標準（practical criteria or actual criteria）兩類，前者係指同儕審查主事機構所訂定之評審標準，或是以調查或訪談同儕審查的利益關係者所得到之評審標準；實際標準則是在同儕審查的過程中，期刊編輯、獨立評審者或評審委員等在實際評審作業時所使用的標準。

六、期刊同儕審查（journal peer review or editorial peer review）

學術期刊針對投寄的稿件所採行之同儕審查作業，期刊同儕審查的評審過程通常分為兩個階段，首先由期刊編輯就期刊發行之目的及主題優先性進行初審，再將通過的稿件選擇適當的 2 至 3 位領域專家進行科學品質審查，並提出接受或

doi:10.6342/NTU201602255

拒絕之出版建議。至於稿件最終的出版命運，則是由期刊編輯參考諸位評審者的意見決定（Campanario, 1998a; Sense about Science [SAS], 2005; Weller, 2002）。

七、獎助同儕審查（grant and fellowship peer review）

政府或民間之學術研究獎助機構針對獎助計畫或獎學金的申請案件所採行之同儕審查作業，獎助同儕審查的評審過程多採兩階段制，先經內部或外部評審者進行獨立審查後，再由內部評審者與獎助機構重要幹部（或董事會成員）所組成的評審委員會進行會議討論，決定最後的獲獎名單（Abdoul, Perrey, Amiel, et al., 2012; Bornmann & Daniel, 2005; Cicchetti, 1991; Marsh & Bazeley, 1999; RIN, 2010）。通常獎助機構會事先訂定評審的規範標準，提供內、外部評審者及評審委員參考（Abdoul, Perrey, Amiel, et al., 2012; Geisler, 2000）。

八、課責性（accountability）

學術界對於課責性的定義看法不一，本研究係指獲得授權之同儕審查的評審者對於授權者及受評者應盡的職責。以期刊同儕審查為例，有些期刊編輯及學者認為評審者是一個同時具有特權與責任的工作，除了專業知識外，評審者做為服務期刊與作者的角色，應具有負責任的態度與適當的禮節，Drotar（2009）認為評審者的評審報告應具有清楚性、專指性、建設性，以及完整性。

doi:10.6342/NTU201602255

第二章　文獻分析

本研究目的除了整體回顧同儕審查的起源、研究現況與展望外，並以我國社會暨人文科學期刊之同儕審查的評審報告進行實徵研究。文獻分析之第一節及第二節說明同儕審查的定義、分類與優缺點，並追溯同儕審查的起源與發展；第三節概述同儕審查的研究現況與國際合作；第四節針對同儕審查的評審標準、信度與公平性的進行討論；最後一節介紹目前各界對於提升期刊同儕審查評審品質所做的努力。

第一節　同儕審查的定義、分類與優缺點

在人類的各項活動中，科學研究可能是受到最多檢驗與評鑑的項目之一（Laloë & Mosseri, 2009），以確保學術研究品質，並做為有限學術資源的分配機制（Bornmann & Daniel, 2004; Cole & Cole, 1973; Frodeman & Briggle, 2012; Lamont, 2009; Langfeldt, 2006）。但是同儕審查有其先天限制，除了評審是人類的行為，易於受到固有人性弱點與個人偏見的影響外，同儕審查機制的設計，讓評審者擁有接近絕對的權力，而且過程既不公開且不透明，因而也引發眾多的質疑與批評（Frishauf, 2009; Geisler, 2000; Wenneras & Wold, 1997; Ziman, 2000），一些學者甚至認為繼續採行同儕審查的唯一理由是缺乏其他更好的方法（Eisenhart, 2002; Harnad, 2008; Kassirer & Campion, 1994; Young, 2003）。

一、同儕審查的定義與分類

基本上，同儕審查是一個通稱，在不同的機構有不同的作法，甚至同一機構的不同單位亦可能有所差異（GAO, 1999; OECD, 2011a; Wood & Wessely, 2003）。社會心理學家將同儕審查的過程概念化，稱之為「在小團體中，針對個人進行社

doi:10.6342/NTU201602255

會評價的過程，例如期刊編輯或評審者審查稿件。」（Bornmann, 2011b, p. 200）有些學者將同儕審查類比為農產市場的穀物分級技術，不過同儕審查的分級並非單一程序，而是一組有彈性且可調整的機制（Chubin & Hackett, 2003; Cole & Cole, 1973; Kostoff, 2004; RCUK, 2006; RIN, 2010）。

以下列舉兩則學術界經常引用的同儕審查定義，一者就運作層面來看：「Michael Gibbons 與 Luke Georghiou 認為同儕審查是某一特定領域的專家，對於相同或相近領域之其他科學家的研究作品，進行科學特定面向（如研究品質）的專業價值判斷。其前提係基於評審專家必須對領域的認知發展、研究方向，以及研究社群有充足的知識。」（OECD, 2011a, p. 1）一者強調目的與應用：「同儕審查是一套科學研究的品質評價機制，科學界利用同儕審查來確認研究程序的正確性及推論的合理性，並根據評審結果分配有限資源，例如期刊版面、獎助名額、學術聲望與特殊榮譽等」（Chubin & Hackett, 1990, pp. 1-2）。

同儕審查原本只是學術界的一種封閉、自成體系的評鑑過程，科學家利用這個機制將學術研究品質的控管權力限縮在科學社群之內，形成所謂的科學自治（Polanyi, 1962）。直到 20 世紀後期，各國政府強調政策目標管考，開始重視公共獎助機構的課責性，使得科學研究不再只是學術界的獨立智識活動，亦需顧及廣大社會的需求（Cozzens, 2001; Frodeman et al., 2012; Hackett, 1997）。為此，美國能源部能源效用與再生能源辦公室（Office of Energy Efficiency and Renewable Energy）也將同儕審查重新定義為：「合格且獨立的評審者，依據客觀標準，以嚴格、正式且做出書面紀錄的程序，就某一研究計畫的科技與商業價值、真實或預期效果，以及生產與管理效用進行評審」（Office of Energy Efficiency and Renewable Energy, 2004, p. v）。OECD 創新政策平台的同儕審查專刊，將這種新型態作法稱之為延伸同儕審查，除了有較多元的評審標準外，例如策略、執行效能，以及科學價值外，而且也會邀請非領域專家擔任評審（OECD, 2011a）。

同儕審查在學術界的應用範圍甚廣，分類方式亦不相同。英國人文社會科學

doi:10.6342/NTU201602255

院（The British Academy, BA）將同儕審查分為兩個層級，第一級同儕審查是直接針對科學家的研究產出進行評審，例如期刊同儕審查及獎助同儕審查等。而若只是在評審過程中參考第一級同儕審查的結果，則屬於第二級同儕審查，例如高等教育機構評鑑或世界大學評鑑等，近年來第二級同儕審查的應用已有益形增加的趨勢（The British Academy [BA], 2007）。Frodeman 等人（2012）則依據評審的用途將同儕審查區分為三類，其一為預期性用途，期刊同儕審查及獎助同儕審查均屬此類，評審者不但要審查期刊稿件或獎助計畫的科學品質，還須顧及評審結果的預期效度；其二為反省性用途，主要檢驗學術領域所採行的某種制度是否達到預訂目標，例如期刊品質評鑑或學術機構評鑑等；其三為綜合性用途，結合預期性及反省性兩種用途，例如大學教職的聘用／升遷同儕審查，評審者除了回溯受評者的過去成就，也必須預測其未來的表現。

比較普遍採用的分類法是以評審標的物為基準，研究者綜整為三類：（一）出版品同儕審查：包括文獻類出版品（如期刊文獻及專書等）及非文獻類出版品（如紀錄片、資料庫、網站，以及電腦軟體等）的同儕審查；（二）獎助同儕審查：包括獎助計畫及獎學金的同儕審查；（三）成就同儕審查：評鑑個人、團隊、部門或機構的學術表現，並依據評審結果分配資源或獎勵，包括大學教職聘用／升遷同儕審查、學術榮譽或獎項同儕審查，以及教育暨研究機構評鑑等（Harley & Acord, 2011; POST, 2002; RIN, 2010）。

二、同儕審查的優缺點

理論上同儕審查的優點似乎不言自明，科學家的同儕當然是最能夠對其研究提出精確看法與建議的人（Chubin & Hackett, 1990）。支持者認為同儕審查的效果至少比科學家的自我約束為佳，並且可提供建設性的理性評價以改進研究品質（Bornmann & Daniel, 2005; Nickerson, 2005）。而若就宏觀層面的科學發展來看，科學研究必須經過批判與檢驗，才得以降低錯誤，成為真正的科學知識（Popper,

doi:10.6342/NTU201602255

1961; Ziman, 2000）。

許多調查研究顯示，學術界對於同儕審查廣泛接受且高度支持（Bertout & Schneider, 2004; Boden et al., 1990; Fletcher & Fletcher, 2003; Gibson, Spong, Simonsen, Martin, & Scott, 2008; POST, 2002; Royal Society, 1995; Ware & Monkman, 2008）。Geisler（2000）整體論述同儕審查的優點，包括提供稱職專家意見、減少低品質科研、管控科研質量、平衡不同科學觀點或思想學派、採取理性、有效與公平的過程、承擔科學發展的責任，以及協助科技資源分配與決策。Brown（2004）及 Ware（2013）則指出期刊同儕審查的多項功能，例如文獻的註冊、保存與改進品質、學術知識的過濾與淨化、學術領域範圍的形塑與發展，以及評審者個人知識的提升等。

不過同儕審查也招致許多批評，研究者綜整如下：（一）信度：評審者之間的一致性過低，可能存有潛在偏見；（二）公平性：評審過程受到非科學價值因素的影響，例如受評者的特殊背景或評審者的個人偏見等，而導致評審不公的情況；（三）效度：評審的效度證實不易，也缺乏客觀指標；（四）效率：評審作業耗費時日且所費不貲；（五）創新性：評審結果傾向於保守主義，未能鼓勵創新；（六）累積性：評審方式比較有利於著名或資深學者，不利於缺乏學術成就的年輕科學家。其他的批評還包括評審過程不透明、評審者剽竊研究創意，以及研究內容偽造或抄襲等，都是經常被討論的問題（Abbott, 2008; Bornmann, 2011b; Bornmann & Daniel, 2005; Braben, 2004; Chubin & Hackett, 1990; Gillet, 1993; Gluckman, 2012; Horrobin, 1990, 1996; Langfeldt & Kyvik, 2011; Luukkonen, 2012; RCUK, 2007; Rip, 2000; Roy, 1985; Travis & Collins, 1991）。

總之，同儕審查是學術社群的自我規範機制，並已緊密地納入學術組織的運作之中，許多學者認為同儕審查有其先天限制，研究者綜整提出同儕審查的八大特質，這些都是在規劃同儕審查作業時必須面對的議題。

（一）「同儕」的定義不易：根據同儕審查的定義，學術領域是決定評審者的主要

doi:10.6342/NTU201602255

關鍵，但是學術領域的界線本有爭議（Price, 1963），而且即使是相同學門的子領域之間，其專業性也未必可以彼此互通。近年來科學專業性的發展益加多元，愈來愈多跨領域、超學科，以及團隊科學的研究，使得同儕的界定更為困難（BA, 2007; Hicks & Katz, 1996; Holbrook, 2013a, 2013b; Lamont, 2009; Langfeldt, 2001, 2006; Wood & Wessely, 2003），個別領域專家在評審跨領域研究時，極有可能面臨專業知識不足的問題（Bornmann, 2011b; Frodeman et al., 2012; Holbrook, 2013a; Huutoniemi, Klein, Bruun, & Hukkinen, 2010; OECD, 2011a）。

（二）「同儕」的角色矛盾：同儕審查最顯著的特徵之一是評審的主客體有可能互換，今天擔任評審者的專家，明天可能在另一場合與現在的受評者直接競爭，這種既是裁判又是球員的情況，很難避免利益衝突（Abdoul, Perrey, Tubach, et al., 2012; Langfeldt, 2001）。尤其在某些研究者人數較少的高度專業化領域，或者小規模科研國家，這個問題就更加複雜（Chubin & Hackett, 1990; OECD, 2011a; Pouris, 1988）。此一潛在風險也促使愛爾蘭及以色列等國家，在進行學術相關評鑑時只邀請外國評審者參與（Gluckman, 2012）。

（三）「評審標準」的解讀彈性：以期刊同儕審查為例，評審者除須判斷稿件的重要性與合理性外，還須對整體研究提出支持或否決的建議，這些都是具有高度主觀性的議題，每個人的看法不盡相同。目前同儕審查的主事機構大都在事前訂定評審標準或原則，但是評審者在實際操作時，往往對於既定的標準有不同的解讀或權重，因而造成評審者之間一致性過低的現象，也引發各界對評審不公的質疑（Bornmann, 2011b; Chubin & Hackett, 1990, 2003; Cicchetti, 1997; Kostoff, 2004; Rennie, 2003; Weiser, 2012; Ziman, 2000）。

（四）「科學品質」的求同求異：Polanyi（1962）指出科學的專業標準包括信度、科學價值，以及原創性，前二者主要強調科學研究應有其一致性，但是原

doi:10.6342/NTU201602255

創性則是鼓勵異議與批判，此兩項特色在本質上就有矛盾之處。同儕審查作為科學研究的評審機制，自然也會面臨相同的難題，除了要評量出具有一致性的核心研究外，還要能挑選出各個領域中最具潛力的創新想法，此一困境由各領域學者對於研究前沿（research front）的低度共識即可得證（Cole, 2000; Wood & Wessely, 2003）。

（五）「多重目標」的取捨與妥協：同儕審查作業所涉及之利益關係者眾多，主事機構往往訂有多重目標，但是各個目標之間卻經常存在彼此消長的關係。以獎助同儕審查為例，獎助機構的目標通常包括效率、效用、課責性、回應性、穩定性、合理性、創新性，以及公平性等，但是若強調了課責性就必須犧牲一些創新性、提升穩定性將會降低回應性，而堅持了公平性則可能要以效用做為代價（Chubin, 1994; Chubin & Hackett, 1990; Luukkonen, 2012; RCUK, 2006; Wood & Wessely, 2003）。

（六）少數人決定的「閉門民主」：同儕審查經常被批評為黑箱作業，主要原因有二，其一、同儕審查是由少數人挑選的少數評審者進行審查，評審報告在本質上雖然只是評審者的個人意見，但是在實際運作時卻代表整體科學家的決定（Harnad, 1996; Perper, 1989; Ziman, 2000）；其二、同儕審查作業強調機密性，在過程不公開且不透明的情況下，容易受到評審者個人價值觀或自身利益的影響，而造成評審結果不公（Chubin & Hackett, 1990; The Higher Education Academy, 2009; Rennie, 2003）。

（七）「自治與責任」的挑戰：20 世紀末期以來，各國政府要求學術領域在科學自治外，也要回應民眾的需求，兼顧科學研究的社會責任。現今全球許多國家的政府獎助機構已直接或間接地將社會影響（social impact）或國家利益納入評審考量，以爭取民眾對政府科研預算的支持。為此同儕審查的自治性與獨立性正在弱化，而逐漸成為一種服務國家政策目標的機制，這種科學認識論與政治問題結合的用途，更加強化了同儕審查作業的矛盾與兩難

doi:10.6342/NTU201602255

（Bhattacharya, 2012; Bornmann, 2013b; Collins & Tabak, 2014; Cozzens, 1999; Frodeman et al., 2012; Holbrook & Frodeman, 2011; Holbrook & Hrotic, 2013; Kamenetzky, 2012; Rip, 2000）。

（八）「信賴與監督」的平衡：同儕審查的正當性是基於學術社群成員之間的彼此信賴與誠信（Abdoul, Perrey, Tubach, et al., 2012; Giraudeau, Leyrat, Le Gouge, Léger, & Caille, 2011; Williamson, 2003），有些學者認為 Merton（1942）的科學四大規範（普遍主義、公有主義、無私利性，以及有條理懷疑主義）足以制約科學家的評審行為（Frodeman et al., 2012; Heitman, 2002），但是許多研究已經證實評審者存有各種偏見，並呼籲建置同儕審查的檢驗與監督機制，以提升評審的公平性（Biagioli, 2002; Callaham, 2003; De Vries et al., 2009; Gluckman, 2012; Levy, 1984; Smith, 2003; van Rooyen, Black, & Godlee, 1999）。

第二節　同儕審查的起源與發展

同儕審查是學術領域的核心，在期刊稿件出版、獎助資源分配，以及大學教職的聘用與升遷上都扮演著關鍵角色。不過就同儕審查的起源來看，除了期刊同儕審查有其歷史成因，可溯源自 17 世紀英國皇家學會出版的《哲學學報》（Biagioli, 2002; Kronick, 1990; Lock, 1985; Zuckerman & Merton, 1971）。獎助計畫與大學教職聘用／升遷的同儕審查則是在 19 至 20 世紀間因應實際作業需要而獨立發展的機制，今天研究獎助機構及大學已廣泛採用同儕審查機制（Bornmann, 2011b; Weiser, 2012）。近年來同儕審查的應用範圍不斷擴大，逐漸成為各國政府政策制訂、法令規章增修，以及學術研究機構考評等作業的重要工具（Frodeman et al., 2012; Guston, 2003）。

doi:10.6342/NTU201602255

一、期刊同儕審查的起源與發展

（一）期刊同儕審查的早期史

1、出版檢查與稿件品質控管

學術界採用的同儕審查機制究竟起於何時，學者看法不一。Kronick（1990）認為就廣義來看，同儕審查遠在人類開始探索與傳播新知時，就已經存在，因為同儕審查（無論發生於出版前或出版後）是建立共識的必要作法，也是科學知識成長的基礎。而大部分的學者則將同儕審查溯源至 17 世紀英國皇家學會出版的《哲學學報》，該刊於 1665 年創刊，由學會首任秘書 Henry Oldenburg 擔任編輯，每期學報出刊前必須先由學會會員進行內容審查，而被視為期刊同儕審查的濫觴（Burnham, 1990; Kronik, 1990; Lock, 1985; Rennie, 2003; Spier, 2002a; Zuckerman & Merton, 1971）。

Biagioli（2002）認為早期期刊同儕審查與政府的出版檢查制度相關，15 至 16 世紀，印刷術在歐洲快速傳播，歐洲各專制政權為了控制言論，實施出版特許制度，規定任何文字內容均須經由政府核可才得以印刷販售。17 世紀中葉，科學研究逐漸機構化，歐洲各地由皇家支持的科學學會興起，並授予出版特許以推廣科學新知，此一作法等同將科學文獻的內容檢查工作交由科學學會負責。整體來看，當時科學學會的期刊同儕審查屬於國家出版檢查體系的一環，各學會除了評審稿件的科學品質外，也同時進行內容檢查，以排除與政府當局不一致的論點（Frodeman et al., 2012; Zuckerman & Merton, 1971）。

到了 18 世紀初期，學術期刊出版市場逐漸成形，一方面因為歐洲各國科學學會積極出版優良期刊，以建立聲望並鞏固地位；另一方面是科學家努力發表期刊文獻，以取得進入科學社群的資格，因此期刊同儕審查的稿件品質控管功能逐漸超越了內容檢查。此一發展也促使期刊同儕審查在歐洲專制政體垮台之後，得以擺脫出版檢查幫手的負面形象，由政府出版檢查制度的執

doi:10.6342/NTU201602255

行者，轉身成為期刊稿件品質的管理者（Biagioli, 2002; Lock, 1985）。

然而直到 19 世紀末期，學術期刊並未立即全面採用同儕審查，主觀因素在於期刊編輯或出版者憂心編輯大權旁落，不願意借重外界專家協助審核文稿；客觀因素則是期刊的數量快速成長，稿件不足的情況甚為嚴重，當時期刊編輯的首要任務是尋找好文章來填補版面，必要時才將稿件交由專家審查（Burnham, 1990; Rennie, 2003; Spier, 2002a）。

2、學科專門化與稿件供需逆轉

期刊普遍採用同儕審查主要出現在 20 世紀，原因之一是期刊稿件供需情勢逆轉，隨著學科專門化以及科學家的人數愈來愈多，研究文獻快速成長，使得期刊稿件不足的情況消失了，編輯工作的最大難題不再是尋找稿件來填補版面，而是如何在眾多投稿中挑選出最佳作品。此外在日益走向專業化的年代，期刊編輯必須考量評審的客觀性以及學者專家的需求，因此有些期刊開始以內聘或委外方式將稿件交由專家評審（Burnham, 1990; Rennie, 2003）。

儘管如此，同儕審查機制並未立即成為顯學，各個期刊採行同儕審查的理由分歧、起始時間不一，評審的作業方式也多有差異。Burnham（1990）以生醫期刊進行研究發現，各期刊開始實施同儕審查的年代沒有顯著的高峰期，身為期刊編輯或發行人的醫生或科學家，大都習於獨自面對專業學門上的變遷或挑戰，也很少參考或直接援引其他期刊編輯的作法。直到二次大戰後期，愈來愈多期刊編輯將稿件交由專家評審，期刊市場採用同儕審查機制的趨勢已然成形，不過同儕審查仍然沒有標準程序，而且各個期刊有不同的作業方式（Manske, 1997）。

（二）期刊同儕審查的當代發展

當代期刊同儕審查普及化的另一股助力來自於書目計量（bibliometrics），20 世紀中葉，Garfield 利用引文分析提出期刊影響係數（Impact Factor）（Garfield &

doi:10.6342/NTU201602255

Sher, 1963），使得期刊同儕審查的效度具象化，同儕審查不再只是為期刊文獻的品質背書，也具有期刊品牌保證的功用。許多期刊資料庫亦將同儕審查列為期刊收錄的必備條件之一（OECD, 2011a），間接地促進期刊同儕審查的發展。

到了 20 世紀後期，書目計量的各項指標，尤其是引文分析，又以其客觀、公平及簡易的優勢，逐漸成為評鑑個人或機構研究表現的重要參考，促使學者積極在同儕審查期刊發表文獻，除了累積個人學術成就外，並可提升獲得研究獎助或擔任大學教職的機會。因此期刊同儕審查再次經由書目計量的連接，間接成為影響獎助計畫申請及大學教職聘用與升遷的重要因素（Harley & Acord, 2011; Harley et al., 2010; van Arensbergen et al., 2014a）。另外對於身處資訊爆炸時代的期刊讀者來說，期刊同儕審查的文獻品管功能，也是值得信賴的閱讀保證。根據英國非政府組織 Sense about Science（2010）針對期刊文獻作者及評審者進行的全球性調查發現，84%的受調者認為若無期刊同儕審查機制，學術傳播將失去控制。

經過 300 多年的發展，期刊同儕審查已成為科學知識的守門者，受到學術社群的重視，但是也有批評與改革的聲浪。近年來網路科技的發展使得期刊同儕審查出現許多創新作法，例如預印同儕審查的 arXiv 典藏庫及出版後同儕審查的 PLoS 電子期刊，都是相當成功的例子；後者除可讓讀者留言評論外，每篇文獻亦提供閱讀、引用及下載次數等最新統計。但是這些成功並不代表傳統期刊同儕審查即將消失，尤其在各國政府與學術界對書目計量益加重視的情況下，學者仍然必須努力在同儕審查期刊發表研究成果，以提升個人的學術成就與聲望（Harley & Acord, 2011; Rennie, 2003）。

今天期刊同儕審查所面對的是一個快速轉換的年代，電子期刊出版的發展及品質持續改進，加上人類閱讀習慣的改變，許多學者認為期刊同儕審查在不久的將來可能發生革命性的變化。Smith（2003, 2009）認為期刊同儕審查的缺點大於優點，在過去數百年之所以停滯不前，主要是缺乏競爭者，他相信未來的同儕審查作業將會更加公開與透明，而企業界的作業流程再造及持續改進等管理作法，可

doi:10.6342/NTU201602255

能全面提升同儕審查的效率、效用與公平性。

二、獎助同儕審查的起源與發展

（一）獎助同儕審查的起源－以英美兩國為例

除了學術期刊外，研究計畫獎助機構是最廣泛採用同儕審查的單位（Bornmann, 2011b）。19 世紀以來，英美等國陸續出現支持科學研究的機構，當時即經常由科學家組成的委員會協助分配獎項，例如 1831 年成立的英國科學促進會（British Association for the Advancement of Science），下設各種不同領域的專業委員會，負責評審各類研究申請案件。美國華府卡內基研究所（Carnegie Institution of Washington）亦於 1902 年設置了 18 個主題諮詢委員會，評審並推薦獎助申請案件；而美國國家研究委員會（National Research Council）在 1919 年就利用同儕審查機制決定獎學金人選（Burnham et al., 1987）。

基本上獎助同儕審查的發展與政府職能擴張息息相關，當各國政府開始承擔支持科學研究的責任，同儕審查就逐漸成為政府獎助經費的主要分配工具（Frodeman et al., 2012）。以英國為例，第一次大戰後期 Richard Haldane（1856–1928）授命檢討政府效能，所提建議之一是將政府資助的科研計畫分為指定及非指定研究兩類；前者由政府主導獎助決策，後者則設置獨立的科學研究委員會，讓科學家共同決定研究的優先順序與經費分配，以強化科學發展的自主性，並避免來自政治人物或行政部門的干預，這就是著名的霍爾丹原則（Haldane Principle）。兩年之後（1920 年）英國醫學研究委員會（Medical Research Council）成立，採用同儕審查方式分配醫學類非指定研究獎助經費（Boden et al., 1990; RCUK, 2006; RIN, 2010）。目前英國有七個不同領域的研究委員會，都是依據皇家憲章設置的準政府機構，由政府科技主管部門訂定各個委員會的經費預算與政策目標，並考核其運作績效。根據英國醫學研究慈善協會（Association of Medical Research Charities）統計，英國各種公共基金每年發放出去的醫學研究獎助經費有 20 多億英磅，其中

doi:10.6342/NTU201602255

超過 95%是透過同儕審查機制進行分配（Ismail et al., 2009）。

美國獎助同儕審查的發展主要在二次大戰之後，政府開始編列巨額預算獎助科學研究。與英國的發展相較，美國的獎助同儕審查在一開始就有兩種不同作法，一為法制化模式，另一為強勢計畫管理者模式（Chubin & Hackett, 1990; Guston, 2003）；前者以美國國家衛生院（National Institutes of Health, NIH）為代表，聯邦政府以法令規定 NIH 及其所屬機構必須採用同儕審查機制；後者主要起源於美國海軍研究署（US Office of Naval Research, ONR），該署在 1940 年代末期開始設立科學研究獎助基金，讓獎助計畫管理者擁有最後決定實權，而同儕審查機制則非必要選項（Burnham et al., 1987; Frodeman et al., 2012）。1950 年美國國家科學基金會（National Science Foundation, NSF）成立，仿效 ONR 的強勢計畫管理者模式，但是也私下納入 NIH 的同儕審查作法（England, 1982; Frodeman et al., 2012）。

20 世紀後期，美國政府強調施政績效管理，美國科技政策辦公室（Office of Science and Technology Policy）為了促進聯邦獎助機構採用同儕審查機制，自 1996 年起聯合管理與預算辦公室（Office of Management and Budget）共同提出年度指導綱領，要求各聯邦獎助機構增加同儕審查評選之獎助比例（GAO, 1999）。到了 2000 年，美國聯邦 R&D 預算中共有 260 億美元（31.4%）係以同儕審查機制進行分配（Guston, 2003）。另外美國政府獎助機構也受到較多政治團體的影響，美國國會主張的學術指定撥款（earmark）即為一例（Savage, 1999; Scarpa, 2009）；有些國會議員認為由科學家主導的獎助同儕審查存有地理偏見，要求直接依據地理區域分配獎助經費。根據美國科學促進會（American Association for the Advancement of Science, AAAS）統計，2010 年美國國會 R&D 學術指定撥款約 42.7 億美元，占 R&D 總費用的 2.8%（American Association for the Advancement of Science, 2010）。

（二）獎助同儕審查的當前重要議題

同儕審查原本是學術界的一種封閉、自成體系的評審過程，科學家利用這個機制將學術研究品質的控管權力限縮在科學家族之內，形成所謂的科學自治。1980

doi:10.6342/NTU201602255

年代，英美等國面對龐大的預算赤字，開始檢討政府獎勵科學研究的成效與課責性，要求政府獎助機構的經費運用除了考量科學發展外，還須提出有益社會的證據，為此社會影響指標逐漸成為各國獎助同儕審查的重要評審項目之一（Cozzens, 2001; Frodeman & Briggle, 2012; Hackett, 1997; Kamenetzky, 2012; Martin, 2011; OECD, 2011a; Roberts, 2009）。

美國 NSF 是較早納入社會影響指標的機構之一，1997 年 NSF 將原來的四項評審標準簡化為智識價值（intellectual merit）與更廣泛影響（broader impact）兩項（Holbrook, 2012; Mervis, 2011）；而歐盟第 5 架構計畫（1998–2002）所採用的 5 項評審指標中，有 3 項與社會影響相關（Holbrook & Frodeman, 2011）。除了前兩者外，Holbrook（2010）認為美國 NIH 及 National Oceanic and Atmospheric Administration、加拿大 Natural Sciences and Engineering Research Council，以及荷蘭 Technology Foundation 也都有類似的做法；而最近啟動的歐盟第 8 架構計畫（Horizon 2020），影響力（impact）仍然是重要審查項目之一。

除了社會影響指標，創新性亦是近年來獎助同儕審查的重要議題。許多文獻指出同儕審查具有保守傾向，不利創新計畫評選（Braben, 2004, 2011; Chubin & Hackett, 1990; Gillett, 1993; Horrobin, 1990, 1996; Langfeldt, 2006; Langfeldt & Kyvik, 2011; Luukkonen, 2012; OECD, 2011a; Roy, 1985; Yalow, 1982; Ziman, 2000）。為了改進此一現象，有些獎助機構調整同儕審查的運作方式，例如要求獎助申請者詳列研究的創新性、聘請非專業人士進入評審團隊，以及強化計畫管理者的決策權限等；有些獎助機構則創設特定獎助項目，例如指定研究主題獎項、創意科學家個人獎項、跨領域或團隊科學研究獎項等，以鼓勵創新研究（Frodeman & Holbrook, 2012; Guthrie, Guérin, Wu, Ismail, & Wooding, 2013; Ismail et al., 2009; Spier, 2002b）。

然而無論是評審社會影響或創新性，都必須面對評審者選聘及評審標準訂定兩項難題（Bell, Shaw, & Boaz, 2011; van der Meulen & Rip, 2000）。學術界除了對

doi:10.6342/NTU201602255

於兩項指標的評審者選聘缺乏共識外，反對社會影響指標的學者認為，考量社會需要可能對科學自治形成干擾，而且社會影響的效用非可立竿見影，也難以推論因果（Bhattacharya, 2012; Bornmann, 2012; Bornmann & Marx, 2014; Bozeman & Boardman, 2009; Holbrook & Frodeman, 2011; Martin, 2011; Rymer, 2011; van der Meulen & Rip, 2000）；而批評創新性指標的學者則表示，所謂創新計畫的定義不夠明確，評審標準各異，而且缺乏可靠的效用考評機制（Donovan, 2011; Frodeman & Holbrook, 2012; Lal & Peña, 2013）。

三、大學教職聘用／升遷同儕審查的起源與發展

（一）大學教職聘用／升遷的制度化－以美國為例

對於學術界人士來說，大學教職聘用／升遷同儕審查是個神祕又令人敬畏的作業，除了評審過程強調機密性外，未能獲得終身教職的老師極可能面臨失業的困境（Weiser, 2012）。相較於期刊稿件及獎助計畫評審，美國大學教職聘用／升遷同儕審查的發展較遲，主要以學術界的同儕審查為根本，並奠基於 1940 年以來逐漸成熟的大學教職職級（academic ranks）與終身教職（tenure）制度（Weiser, 2012）。

早期美國大學教職聘用與升遷的權力大多集中在學校行政部門，教職工作缺乏保障，甚至可能因為學校捐款人或董事會成員向校方施壓而失去工作，直到大學實施終身教職制度才改變此一情況（Brown & Kurland, 1990; Cameron, 2010; Metzger, 1990; Weiser, 2012）。1915 年美國大學教授協會（American Association of University Professors, AAUP）成立，除鼓吹學術自由外，也提出教職解雇作業準則，以及倡議終身教職制度（American Association of University Professors [AAUP], 1915）。之後 AAUP 在 1940 年提出學術自由與終身教職聲明，除詳列由試用到終身教職的進程外，並強調終身教職制度是學術自由的基礎，讓教師擁有研究出版、課堂講授與討論，以及言論發表與寫作的自主空間。但是該聲明並未提及教職聘用與終身教職的評審標準，比較相關的文句是：學術自由的目標是教學與研究

doi:10.6342/NTU201602255

（AAUP, 1940; Weiser, 2012）。

經過 AAUP 多年的努力，目前美國大學普遍採用終身教職制度，教職工作獲得制度性的保障。不過每所學校之教職聘用／升聘同儕審查的作業方式差異甚大，甚至在同一所大學內的不同系所也可能有別，以評審者為例，有的以內部評審為主、有的強調外部評審，有的則是內外部評審兼具（Frodeman et al., 2012; Gross-Schaefer et al., 2015; Weiser, 2012）。1966 年，AAUP 提出大學院校治理聲明，呼籲由大學校務董事成員、行政管理者、教職員、學生及其他成員等共同承擔學校治理的責任（AAUP, 1966）。這股大學共同治理風潮，不但擴增了大學教職聘用／升遷同儕審查作業所涉及之利益關係者，也使得評審的決策權力結構更加複雜多元（Cummings & Finkelstein, 2012; Weiser, 2012）。

（二）全球大學治理的變革

1960 年代，由美國興起的大學共同治理風潮，除了增加教職聘用／升遷同儕審查作業的利益關係者外，也改變了評審決策的權力結構。美國 AAUP 在 1966 年提出大學治理聲明，呼籲由學校的董事會成員、行政管理者、教師、學生及其他人員，共同承擔學校治理的責任（AAUP, 1966）。此一聲明讓大學教師在校務管理上有了合法的地位，包括參與教職聘用與升遷的人事作業，這個趨勢也逐漸擴及歐洲及亞洲國家，形成一股全球性的大學共同治理風潮（黃宗貴，2010a，2010b；Cumming & Finkelstein, 2012）。我國教育部在 1994 年經過立法院通過修正《大學法》，放寬大學自主治理的權限，尤其是 2001 年的校務基金制度採行之後，對於減少教育部對大學營運的約制、保障學術自由，以及建立校內民主機制等具有重大意義（成群豪，2006）。

20 世紀末期，西方國家的高等教育機構因為政府教育預算緊縮而有了新的發展，各國大學及學院開始調整內部組織架構與共同治理的模式。以英國為例，1988 年通過的教育改革法，使得大學教職職位的穩定性受到影響（Becher & Trowler, 2001）；而在 1992 年由政府推動的科技學院升格方案，更加限縮了高等教育的自

doi:10.6342/NTU201602255

主空間，批評者甚至認為大學自我管理的時代已經結束（Trow, 1992）。在美國的情況亦同，政府對大學生的人均投資越來越少，對校務的干預卻越來越多，除了各州政府紛紛採取各種教學品質的管控政策外，1992 年國會更要求各州設置高等教育評鑑組織，邁入了政府管理高等教育的年代（Dill & Sporn, 1995）。然而在英美兩國逐漸強化政府干預大學自治之同時，歐洲大陸國家卻採取權力下放政策，強調大學自我管理的責任與效率，許多大學開始重視校長人選及行政單位效能，積極對外爭取經費或資源，而逐漸走向市場機制經營（黃宗貴，2010b）。

總而言之，高等教育是一個不斷變動的體系，由學術、國家及市場三股勢力交互作用的結果，學者稱為大學治理的三重螺旋。近年來國家與市場的力量透過各種方式影響學術界，使得大學的經營與管理方式產生變革（Becher & Trowler, 2001），這些也都將對大學教職聘用／升遷的同儕審查作業產生重大影響。

（三）大學教職聘用／升遷同儕審查的發展現況

整體來看，大學教職聘用／升遷同儕審查的評審標準係以教學、研究與服務為主，但是自 20 世紀後期以來，許多國家的大學都出現偏重研究的情況（Fairweather, 2005; Greenbank, 2006; Kreber, 2002; Pratt, 1997），原因之一為各國政府高教機構評鑑大多強調書目計量指標；之二是各國政府的高教經費縮減，校務營運益形倚重校外研究獎助資源（Andersen, 2003; Becher & Trowler, 2001; Bornmann, 2013a; Greenbank, 2006; Harley et al., 2010; Waters, 2009）。在這種雙重的壓力下，學者出版期刊文獻及獲得研究獎助之能力，就成為爭取大學教職聘用與升遷的關鍵因素（van Arensbergen et al., 2014a; van Arensbergen, van der Weijden, & van den Besselaar, 2014b）。批評者認為此一趨勢形同將大學教職聘用／升遷同儕審查的工作，委由期刊出版商及研究獎助機構辦理（Boyer, 1997; Harley & Acord, 2011; Harley et al., 2010）。

除了偏重研究外，許多國家的大學教職聘用／升遷同儕審查也有忽視教學表現的情況。英國的大學教師多意識到各級政府及學校比較重視研究成績（Dearing,

doi:10.6342/NTU201602255

1997），並認為教學表現與教職升等之間的關聯性不大（Young, 2006）；澳洲的調查也有類似的發現，儘管大多數的學者（88.2%）支持獎勵優質教學，但是只有 31.4% 的教師認為教學表現有助升遷（Bexley, James, & Arkoudis, 2011）。另外 Milem、Berger 與 Dey（2000）指出在 1972 至 1992 年間，美國四年制大學教師從事研究工作的時數顯有增加，但是在課外與學生進行諮商或互動的機會卻相對減少。

近年來，許多國家已經注意到大學校園存在重視研究而輕忽教學的情況，積極調整高等教育管理政策並鼓勵優質教學，以促進研究與教學的平衡。2010 年 OECD 出版高等教育優質教學回顧報告，調查分析全球 20 國、29 家高等教育機構的 46 項優質教學計畫方案，提供全球各大學參考。該報告並指出各國高等教育機構評鑑以及世界大學排名大多過度重視研究指標，而引發輕忽教學的批評，不過如何評鑑教學品質，也是學術界必須積極面對的議題（Hénard, 2010）。

四、其他類型的同儕審查

同儕審查原本只是科學研究產出的評審機制，主要用於審查期刊稿件與獎助計畫的科學價值，而後逐漸擴及大學教職的聘用與升遷作業。20 世紀後期，一些國家將同儕審查應用於高等教育機構評鑑，並據以分配教育研究經費（OECD, 2011b），例如英國大學科研評鑑（Research Excellence Framework，前身為 Research Assessment Exercise）及澳洲大學研究品質框架（Research Quality Framework），都局部採用同儕審查機制，而所採用的評審標準對於大學管理及國家整體科研發展方向，都有重大的影響（Donovan, 2011; Whitley & Gläser, 2007）。

另外同儕審查還有其他衍生用途，尤其在協助訂定政府公共政策或進行機構考評上扮演著重要角色。美國環境保護局要求環保規章的制定要建立在同儕審查的審議基礎上，而該局對於公開的環境資訊，亦採取同儕審查機制為訊息的品質把關（Frodeman et al., 2012; Sarewitz, 2004）。Guston（2003）整理美國政府部門運用同儕審查的情形，包括獎助機構進行的績效評鑑、法令規章修改與制定之評析，

doi:10.6342/NTU201602255

以及法院專家證人說詞的同儕審查等，他認為同儕審查的應用已經超出學術研究的範疇，而且具有強化政府施政的合理性及提升政府效能的功用。

第三節　同儕審查的研究現況與書目計量

同儕審查是學術界有限資源的分配機制，其操作過程卻具機密性，在公眾領域幾乎沒有任何線索，直到 1980 年代前後才開始有較多理性檢驗（Chubin & Hackett, 1990; Rennie, 2003; Weller, 2002）。目前同儕審查的研究以期刊稿件及獎助計畫為主，大學教職聘用與升遷的探討甚少。近半個世紀以來，除了學術界對於同儕審查研究的廣度與深度逐漸提升外，同儕審查的主事機構亦積極進行國際合作，出版評審作業的基本規範或操作指南，以強化審查的品質與公信力。在此之同時，書目計量也憑藉著客觀與簡易的特質逐漸受到各界的重視，近年來更與同儕審查並列為學術評鑑的兩大主要工具。有些學者認為同儕審查與書目計量各有其優缺點，而且具有彼此互補功能，若能整合運用，應可提升學術評審作業的品質與效用。

一、同儕審查的研究現況

同儕審查的實徵研究以公平性與信度為主，效用的研究較少（Bornmann, 2011a; Weller, 2002; Wessely, 1998），不過許多學者指出大多數文獻的方法論不夠嚴謹、因果推論較為薄弱，而且各篇研究結果的異質性甚大難以通則化，建議採用後設分析法或實驗研究加以補強（Bornmann, 2011b; Bornmann, Nast, et al., 2008; De Vries et al., 2009; Demicheli & Di Pietrantonj, 2007; Jefferson et al., 2007），Marsh 等人（2011）則認為除了二手文獻的後設分析外，大規模一手評審資料的研究亦有其必要性。

doi:10.6342/NTU201602255

（一）打開同儕審查的「黑盒子」

期刊同儕審查的理性檢驗在 1980 年代逐漸受到重視，Rennie（2003）舉出兩項具體事證，其一為 1985 年 The British Medical Journal（BMJ）總編輯 Stephen Lock 在英國奈菲爾基金會（Nuffield Foundation）的支持下出版第一本期刊同儕審查專書—《困難的平衡》（*A Difficult Balance*），內容分析期刊同儕審查的作業細節，並呼籲進行更多的討論與研究；其二是 1989 年首屆同儕審查與生醫期刊研討會在美國芝加哥舉行，之後每四年辦理一次，除了論文發表數量逐屆成長外，研究的品質與複雜度也見提升。在這股研究風潮的影響下，加上期刊市場競爭日烈，許多期刊編輯開始進行同儕審查作業的自我檢驗；而透過愈來愈多的研究，期刊編輯也益加無法專斷決定稿件的命運（Altman, 1996; Rennie, 1998a, 1998b, 2003; Smith & Rennie, 1995）。Rennie（1998b）主張將期刊同儕審查的自我檢驗列為期刊評鑑的標準之一。近年來文獻偽造及抄襲的情況屢屢出現（Fang & Casadevall, 2011; Healey, 2013; Steen, 2011; Steen, Casadevall, & Fang, 2013），期刊同儕審查的自我檢驗再度引起各界的重視（Wager, 2008）。

獎助同儕審查的早期研究亦多屬自我檢驗類型，以美國為例，NIH 及 NSF 在 1970 至 1980 年代即曾分別委託學者進行同儕審查作業的自我檢驗（Burnham et al., 1987; Chubin & Hackett, 1990）。Chubin 與 Hackett（1990）將獎助同儕審查的研究分為三類並說明其特色，其一、獎助機構委託學者之研究：獎助機構指定學者並提供研究資料，因為學者類似自己人，研究結論大多偏向機構立場；其二、獎助機構贊助學者之研究：學者雖然擁有較多探索及表達的自主空間，但是研究資料仍需仰賴獎助機構提供，增加了研究的複雜性，研究結論也比較容易失真；其三、學者自發性地獨立研究：學者不但沒有獲得任何經費與社會支援，而且必須自行設法去接近研究目標群體。直到 20 世紀末期，許多國家的公共獎助機構在政府資訊開放及政策目標管考的壓力下，為了提升同儕審查的作業品質與效能，比較願意將第一手評審檔案提供學者進行研究，有些甚至允許學者訪談評審者或觀察評

doi:10.6342/NTU201602255

審委員會的議事與決策過程（Bornmann & Daniel, 2005; Lamont, 2009; Langfeldt, 2001; Marsh et al., 2008; Wood & Wessely, 2003），使得獎助同儕審查的評審作業逐步邁向開放與透明之途。

相較於期刊及獎助同儕審查，大學教職聘用／升遷同儕審查的作業程序更為複雜，除了利益關係者眾多及評審權力結構多元外（Cummings & Finkelstein, 2012），評審過程亦具機密性，相關研究一向不多（Brooks, 1988; Weiser, 2012）。但是到了 1990 年代前後，許多國家的教育經費縮減及學生人數減少，大學教職的穩定性受到影響，此項強調學者個人表現的審查作業開始受到重視，有關評審標準的討論也日漸增加（Becher & Trowler, 2001; Harley & Acord, 2011; Harley et al., 2010; Weinbach & Randolph, 1984），不過大學教職聘用／升遷同儕審查迄今依然處在密閉的黑盒子之中。

（二）同儕審查的研究現況

1、期刊同儕審查的研究現況

Chubin 與 Hackett（1990）將期刊同儕審查研究分為兩類，一者針對評審作業程序，探討稿件審查的準確性或評審意見的信度與效度等；一者視同儕審查為科學研究的守門者，強調評審的公平性，譴責各種類型的偏見。許多學者指出期刊同儕審查文獻的研究設計差異甚大，有針對編輯、評審者或作者進行調查；有探討評審結果與可能變項（評審者或編輯背景等）之相關性；有以編輯或評審者的審查報告進行內容分析或個案研究，不過各篇文獻大都只著重於同儕審查的某個面向，在研究方法上也各有其優點與不足（Bornmann, 2011b; Chubin & Hackett, 1990; De Vries et al., 2009; Jefferson et al., 2007）。Weller（2002）亦發現同儕審查的研究雖然橫跨各個領域，但是主要集中在醫學、社會科學及心理學，而且各學科的研究重點並不相同，例如醫學領域對於統計方法的評審方式甚為關切、心理學與社會學則強調評審者的偏見，而有關拒絕率的研究通常來自於社會科學領域。

doi:10.6342/NTU201602255

Bornmann（2011b）的回顧文獻指出，最近十餘年的期刊同儕審查實徵研究以信度及公平性為主，許多研究發現評審者之間的信度不高，但是評審者對於拒絕稿件的共識程度高於接受稿件;至於公平性問題雖經有些文獻證實，然而因為研究結果不一，難以通則化。另有少數文獻針對同儕審查的預期效度進行檢驗，探討是否評選出最佳稿件，有些作者發現稿件遭到拒絕之後，另在其他期刊登載的比例甚高，而且不必然是知名度較差的期刊，似乎顯示稿件審查除了基於科學品質外，也受到審查過程所在情境的影響；也有學者利用引用數據作為效度指標，比較接受稿件與拒絕稿件之影響力，大都顯現編輯的決定具有高預期效度，還有研究證實接受稿件的被引用情形也高於拒絕稿件（Bornmann, 2010）。

期刊同儕審查迄今只有少數準實驗或實驗研究，Peters 與 Ceci（1982）將美國 12 家著名期刊 2 至 3 年前出版的文獻各選 1 篇（共 12 篇），經變造作者姓名及服務機構後再投稿原出版期刊，結果只有 3 篇被識破，另外 9 篇中有 8 篇又再次經過同儕審查的程序，不過反而成了拒絕稿件。此篇文獻的研究設計雖然遭到學術道德的批評，但是卻被後續研究者廣泛引用，做為批評期刊同儕審查的重要佐證。另外有些期刊主動進行實驗研究，例如評審者的偵錯能力分析（Baxt, Waeckerle, Berlin, & Callaham, 1998），以及雙盲或簽名評審對評審結果的影響等，都是為了提升同儕審查的作業品質與公平性（Godlee, Gale, & Martyn, 1998; Justice et al., 1998; McNutt, Evans, Fletcher, & Fletcher, 1990; Nylenna, Riis, & Karlsson, 1994; van Rooyen, Godlee, Evans, Smith, & Black, 1998）。

近年來期刊同儕審查的發展受到網路及電子出版的衝擊，有些期刊出版者或研究機構主動嘗試新型態評審作法，例如預印、出版後或公開同儕審查等，辦理的成效不一，有的持續至今，有的半途中止（Bornmann, 2011b; Harnad, 2000; Nature, 2006; RIN, 2011）。不過許多獎助機構積極支持這些創新作法或

doi:10.6342/NTU201602255

相關研究，希望及早掌握網路時代最適當的期刊稿件評審模式（Borgman, 2007; Bornmann & Daniel, 2010; Bornmann, Marx, Schier, Thor, & Daniel, 2010; Ford, 2013; Odlyzko, 1996; van Rooyen, Delamothe, & Evans, 2010）。

2、獎助同儕審查的研究現況

公平性與信度是獎助同儕審查長期受到關注的議題，預期效度的研究則較少，重要研究發現分述如下：（1）公平性：許多文獻雖然已經證實存在評審者偏見，不過大多數的學者在探討公平性時，係將多階段的獎助同儕審查視為一個整體，僅由評審結果分析可能的偏見。此類研究的主要問題在於因果推論薄弱，而且各篇文獻的結論異質性甚大，難以通則化（Bornmann, 2011b; Demicheli & Di Pietrantonj, 2007）；（2）信度：獎助同儕審查評審者之間的信度過低是普遍存在的現象（Cicchetti, 1991; Goldman, 1994; Hodgson, 1997; Oxman et al., 1991），有些學者認為低信度來自於評審者的個人偏見（Eckberg, 1991; Kostoff, 1995; Opthof & Wilde, 2009; Wessely, 1998）；有些學者卻認為評審者的看法不一致，符合科學評審的常態（Chubin & Hackett, 2003; Cole, 1992; Stricker, 1991; Wood & Wessely, 2003）；（3）預期效度：少數文獻利用獲獎及落選計畫的引用數據，探討評審的預期效度，各篇研究結果的反差甚大，沒有系統性結論；不過此類研究有其先天困難，因為落選的獎助計畫大多胎死腹中，難以追蹤其效度（Bornmann, 2011a; Langfeldt, 2006; Mutz, Bornmann, & Daniel, 2015）。

近年來獎助同儕審查研究有了大幅進展，愈來愈多政府獎助機構願意開放評審過程資訊提供研究，學者不再只是針對評審結果進行分析，並可探討各階段的審查作業細節（Bornmann, Leydesdorff, & van den Besselaar, 2010），有學者觀察獎助評審委員會的議事運作及決策過程（Lamont, 2009; Langfeldt, 2001, 2006; Luukkonen, 2012; Olbrecht & Bornmann, 2010）；有學者分析內外部評審者的審查報告及決審會議紀錄（Abdoul, Perrey, Amiel, et al., 2012）；亦有

doi:10.6342/NTU201602255

學者探討不同評審階段的評審者所使用的評審標準，以及各階段評審結果的穩定性與相關性等（Bornmann & Daniel, 2005; Bornmann, Mutz, et al., 2008; van Arensbergen et al., 2014a; van den Besselaar & Leydesdorff, 2009）。目前大部分研究指出，獎助同儕審查的內、外部評審者或評審委員會成員，大多依據獎助機構所預訂的評審標準進行審查，但是每位評審者對於各項標準的看法或重視程度有別，是否因而影響評審結果，仍有待進一步討論（Abdoul, Perrey, Amiel, et al., 2012; Bornmann & Daniel, 2005; Langfeldt, 2001）。

獎助機構也進行少數實驗性研究，最著名者為美國 NSF 的評審者一致性實驗，研究團隊推論有一半的獲獎者是受到隨機因素的影響（Cole, Cole, & Simon, 1981）。近年來 European Molecular Biology Organization 進行同儕審查性別偏見檢測，得到不利於女性申請者的結論（Ledin, Bornmann, Gannon, & Wallon, 2007）；另有澳洲研究委員會（Australian Research Council, ARC）以少數專家審閱所有申請案件，發現除了評審者之間的信度提升外，在評審時間與經費上亦較為經濟（Jayasinghe, Marsh, & Bond, 2006）。

3、大學教職聘用／升遷同儕審查的研究現況

大學教職聘用／升遷同儕審查的研究文獻不多，主要探討評審標準、評審作業程序，以及評審決策權力結構等議題。美國當代語言協會（Modern Language Association, MLA）於 2005 年進行的教職升遷調查指出，美國四年制英語及外語系所在評審終身教職時，認為出版品非常重要者達 75.7%，比 Wilcox（1970）在 40 多年（1966–1967）前的調查高出兩倍強（35.4%）；該調查亦發現數位文獻或專論在評審時比較不受重視，甚至有不予承認的情況（Modern Language Association, 2006）。另有一些文獻證實研究計畫獲獎紀錄也是大學教職聘用／升遷的重要因素，例如丹麥研究委員會 Independent Research 的獲獎者成為正教授的比率為 16%，落選者則只有 9%（Bloch et al., 2014）。歐盟研究委員會（European Research Council）的 Starting Grants 及德

doi:10.6342/NTU201602255

國研究協會（Deutsche Forschungsgemeinschaft, DFG）的 Emmy Noether Programme 的獲獎者，也有類似的情形（Hornbostel, Böhmer, Klingsporn, Neufeld, & von Ins, 2009; Laudel & Glaser, 2012）。

關於大學教職聘用／升遷同儕審查的評審作業程序，Weiser（2012）蒐集美國多所大學的運作方式，綜整提出五大共通點，包括評審項目（新聘或教職職級）、評審者（內部評審或外部評審）、評審者的評審範圍、評審標準，以及評審作業的機密性。他並指出內部評審一般分為系級、院級及校級，各級評審委員會的組成及標準都不相同，至於外部評審則以研究型大學較為普遍。另外有關評審決策權力結構的研究，根據德國 Kassel 大學 International Center for Higher Education Research 進行之跨國調查發現，大學教職的研究、教學與服務評鑑所涉及之利益關係者眾多，包括政府與外部利益團體、校級行政部門、系主任與院長、教師委員會或協會、同事，以及學生等六大類，而且各利益關係者對於教師評鑑結果的影響力也隨著國家與機構而有所不同（Cummings & Finkelstein, 2012）。

（三）同儕審查的國際合作

同儕審查的作業方式在大多數的國家均未以法律規範，而是由主事機構自行決定，或有明訂操作指南，或有依內部慣例進行，各個機構的作法甚為多元（Bornmann, 2011b; GAO, 1999; OECD, 2011a; Rennie, 2003; Weiser, 2012; Weller, 2002; Wood & Wessely, 2003）。不過多年來各國同儕審查主事機構亦積極進行國際交流，彼此分享經驗，以提升同儕審查的品質與公信力。目前期刊及獎助同儕審查都已有跨國性的合作成果，有些學者亦主張透過高等教育機構的跨國合作，共同選定大學教職聘用／升遷同儕審查的基本評審指標（Harley & Acord, 2011; Weiser, 2012）。

期刊同儕審查之跨國合作較早，國際醫學期刊編輯委員會於 1979 年出版《生物醫學期刊投稿統一規範》（*Uniform Requirements for Manuscripts Submitted to*

doi:10.6342/NTU201602255

Biomedical Journals），提供作者及期刊編輯參考。2013 年該委員會大幅修訂《統一規範》並更名為《學術研究的管理、報告、編輯與出版在醫學期刊之建議規範》（*Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals*），詳列作者、評審者、編輯、出版者在同儕審查過程中的職能與責任，並強調同儕審查是科學研究過程中不可或缺的一環（ICMJE, 2013）。委員會除呼籲生醫與健康照護期刊支持採行《建議規範》外，並定期更新內容，最新版已於 2015 年 12 月出版。

獎助同儕審查機構的跨國合作發展較遲，歐洲研究機構負責人協會（European Heads of Research Councils）及歐洲科學基金（European Science Foundation, ESF）有感於歐洲各國獎助類型與同儕審查作業的多樣性，於 2010 年合作推動跨國性調查，並於次年出版《歐洲同儕審查指南》（*European Peer Review Guide*）。全書的第一部分為獎助同儕審查作業概覽，說明獎助的類型與差異以及同儕審查的價值與作業方式等；第二部分則分別論述不同類型的獎助計畫，包括個人與合作研究計畫、職業生涯發展計畫、科學網絡創設與強化計畫，以及區域研究中心與研究基礎設施建置計畫等（European Science Foundation, 2011a, 2011b）。

此外美國 NSF 為了建立獎助同儕審查的核心價值，於 2012 年首度召開全球科學同儕審查高峰會，共有近 50 國（多為 G20 及 OECD 的會員國）的獎助機構代表與會，會中決議建立全球研究委員會（Global Research Council, GRC），並提出科學同儕審查六項原則聲明，包括專家評審（expert assessment）、透明性（transparency）、公平性（impartiality）、適切性（appropriateness）、機密性（confidentiality），以及誠信與道德考量（integrity and ethical considerations）；聲明中亦強調獎助機構作為公共資金的管理者，必須展示它們對於送審之研究計畫具有卓越的評審能力，並確保其研究宗旨符合政策目標；而嚴謹且透明的同儕審查作業，將有助於確保政府的獎助經費應用在最能促進科學發展及解決社會問題的計畫上（Global Research Council, 2012）。之後 GRC 年會曾分別在柏林、北京、東

doi:10.6342/NTU201602255

京舉行。

二、同儕審查與書目計量

學術界除了進行同儕審查的研究外，也積極尋求改進或替代方案。近年來期刊同儕審查受到網路科技及電子出版的影響，已出現許多創新作法而呈現多元發展。至於獎助計畫及大學教職聘用／升遷同儕審查，最常引發討論的是同儕審查與書目計量的競合關係，有些學者認為廣泛且多樣的書目計量指標有助於提升同儕審查的合理性與透明性（Bornmann, 2011a, 2013a; van Raan, 2005）。Geisler（2001）指出同儕審查與書目計量的結合應用，讓同儕審查的決定不再只是評審者的主觀意見，也加入了客觀的量化指標；但是如何將書目計量的量化資訊妥適地納入各種不同且多樣的同儕審查機制之中，是學術界面臨的重大挑戰。

（一）書目計量與科學影響力

Alan Pritchard 在 1960 年代提出書目計量，用統計方法呈現已紀錄的資訊，例如計算專書、文獻、出版品，以及引用的數值（Bornmann, 2013a），其中引用數據已逐漸被視為科學影響力的重要指標之一（Daniel, 2005; Garfield & Welljamsdorof, 1992; Smith, 1981）。Merton（1988）認為文獻引用在知識傳遞與擴散上具有雙重意義，在工具意義上，引用代表可能具有參考價值的資訊；而在系統意義上，引用是科學家在知識的智慧財產資料庫中，對於同儕的研究成果進行認知註記。

許多學者質疑引用的動機並指出引文分析的限制，包括引用行為偏好、引用效用可議，以及引文資料庫的正確性不足等；甚至有學者認為學術傳播系統本不完美，出版品的重要性不等同於影響力，而且被大量引用的文獻也未必一定是高品質（Bornmann, 2011a; Laloë & Mosseri, 2009; Martin & Irvine, 1983; Peters & van Raan, 1994）。不過 Bornmann 與 Daniel（2006）回顧 1960 至 2005 年出版的引用行為實徵研究文獻發現，雖然大多數研究證實引用動機不只是對於科學家同儕之智

doi:10.6342/NTU201602255

識與認知影響表達認同，也受到其他非科學價值因素的影響。但是這些研究被視為缺乏信度，因為各篇的研究設計差異甚大，研究結果幾乎無法複製，而且很多文獻在方法論上有其缺陷。

Van Raan（2005）認為許多研究顯示引用動機並非如此的不同或隨機，反而在許多情況下，引文分析確實是影響力的可靠指標。支持引文分析的學者表示，儘管許多學者認為同儕審查是品質保證的象徵，但是對於大多數的人來說，同儕審查只是提供出版品給科學社群，但是接受度（引用）才是影響力的代表（Bornmann, 2013a; Shadbolt, Brody, Carr, & Harnad, 2006）。Research Evaluation and Policy Project（2005）指出雖然研究品質的評鑑需由同儕進行，但是許多國家的政策已逐漸將影響力視為品質的代理指標。

（二）同儕審查與書目計量的互補性

質性的同儕審查與量化的書目計量是當前學術領域的兩大評審方式，各有其支持者，也受到許多批評。許多文獻探討同儕審查與書目計量的互補關係，有些學者認為學術評鑑應始自學術價值、終於學術影響力，因此同儕審查必須結合書目計量，才得以完整呈現學術評鑑的全貌（Borgman, 2007; Bornmann, 2011a; BA, 2007; Pendlebury, 2008）。Wouters（1997）以資訊流的觀點提出科學知識循環模式，分析同儕審查與書目計量的循環互動關係，認為透過書目計量之檢驗，同儕審查過程中的領域專家將難以專斷自主。

Van Raan（1996）認為書目計量指標不應該單獨使用，建議將強調研究表現的書目計量指標（尤其是引用數據）納入同儕審查的評審過程之中，以作為評審者的重要的參考資訊。此一作法使得同儕審查不再只是少數評審者的意見，評審者可以透過書目計量瞭解受評者在全球研究前沿的地位、影響力與特殊性，並可進一步洞察科學傳播模式與知識散佈過程，此即所謂的「資訊充分的同儕審查」（informed peer review）。Bornmann（2013a）指出書目計量與同儕審查整合有兩大效益，其一、廣泛且重要的書目計量指標，有助提升同儕審查的透明性與合理性；

doi:10.6342/NTU201602255

其二、書目計量可以檢驗同儕審查的結果，以避免學閥派系效應，而此點正是同儕審查的根本問題。他也認為書目計量指標必須經過專家評審者的挑選與解讀，才可以用來評鑑受評者的專業表現。

目前同儕審查與書目計量的整合評鑑方式，已經應用在機構層級的學術評鑑，例如歐洲許多國家採行的高等教育機構評鑑，即同時包括同儕審查與書目計量指標。但是此一方式應用在個人層級的評鑑時，就引發較多爭議，如何選擇適當的書目計量指標、如何兼顧領域的差異性，以及如何分配兩者的權重等，都是受到關切的議題，這些亦是獎助計畫與大學教職聘用／升遷同儕審查必須面對的重要議題（Abramo & D'Angelo, 2011; Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015; Cabezas-Clavijo, Robinson-García, Escabias, & Jiménez-Contreras, 2013; Harley & Acord, 2011; Harley et al., 2010; Weingart, 2005）。

第四節　同儕審查的評審標準、信度與公平性

同儕審查是學術界的品質控管機制，具有結合眾人之智的資訊整合功能，以及在決策團體中的共識建構功能（Nijstad, 2009）。一般來說，參考多人意見而形成的決定，外界的接受度理應較高（Olbrecht & Bornmann, 2010），但是已有研究證實同儕審查評審者之間的信度過低，而且也出現評審不公的情勢，因此對於同儕審查的評審標準、信度，以及公平性，一向是研究同儕審查的學者所關切之重要議題。

一、同儕審查的評審標準研究

理論上，研究評審標準最直接的方式是利用第一手的評審者審查報告，瞭解評審者在審查作業時所使用的實際標準，但是因為涉及評審過程與檔案的機密性，相關研究不多。目前大部分的文獻是針對同儕審查的利益關係者進行之意見調查，

doi:10.6342/NTU201602255

所得結論學者稱之為規範標準（normative criteria 或 theoretical criteria），以與評審者的實際標準（practical criteria 或 actual criteria）有所區隔（Chubin & Hackett, 1990; Weller, 2002）。20 世紀末期，歐洲、美加，以及澳洲等國家的政府獎助機構陸續開放評審檔案提供研究，獎助同儕審查也正式邁向實際標準的研究年代。另外近年來因為大學教職的競爭益烈，有關大學教職聘用／升遷同儕審查所使用的評審標準，也逐漸受到學術界的重視與討論（Harley & Acord, 2011; Harley et al., 2010; Weiser, 2012）。

（一）期刊同儕審查的評審標準

期刊同儕審查的評審過程通常分為兩個階段，首先由期刊編輯就期刊發行之目的及主題優先性進行初審，再將通過的稿件選擇適當的領域專家 2 至 3 位進行學術研究的品質審查，並提出接受或拒絕之出版建議。至於稿件最終的出版命運，則是由期刊編輯參考諸位評審者的意見決定（Campanario, 1998a; SAS, 2005; Weller, 2002）。

目前期刊同儕審查的評審標準文獻大多是針對個別期刊（少數為跨期刊）或小範圍的利益關係者（如作者、評審者、編輯及讀者等）之調查訪問，少數利用評審者的審查報告進行研究。Bornmann、Nast 等人（2008）蒐集 1967 至 2006 年間出版的 46 篇期刊同儕審查之評審標準文獻，利用後設分析法歸納編輯與評審者所採用之 542 項評審標準（若計入重複分類者共 572 項），再將之細分為 9 大類、35 小類，每項標準並區分為正面、負面及中性表述。表 2-4-1 根據 9 大標準在各篇文獻中的出現次數依序排列如下：貢獻度相關、寫作與呈現、設計與概念、方法與統計、結果討論、文獻分析、理論、作者聲望或所屬機構，以及道德。整體而言，期刊編輯與評審者使用的負面表述多於正面，最常使用的 2 項標準為貢獻度相關及寫作與呈現，後者有一半以上為負面評論，可能是提出批評意見時經常採用的標準，至於非關文獻學術品質的作者聲望及所屬機構僅排名第 8。

doi:10.6342/NTU201602255

表 2-4-1 期刊同儕審查的 9 大評審標準—依文獻中出現次數排序

標準／理由	總數／篇數	正面	負面	中立
1. 貢獻度相關	148／45	36	47	65
2. 寫作與呈現	143／44	22	78	43
3. 設計與概念	92／43	19	40	33
4. 方法與統計	72／34	8	30	34
5. 結果討論	45／31	10	16	19
6. 文獻分析	27／26	3	10	14
7. 理論	24／22	5	11	8
8. 作者聲望或所屬機構	11／11	3	2	6
9. 道德[a]	10／13	0	5	5
合計	572／46	106	239	227

a. 含多次或分割出版文獻，所以篇數 13 大於出現總數 10。

資料來源：“Do editors and referees look for signs of scientific misconduct when reviewing manuscripts?” by L. Bornmann, I. Nast, and H.-D. Daniel, 2008, *Scientometrics, 77*(3), pp. 415-432. 研究者節錄製表

為進一步探討期刊編輯及評審者對 9 大標準的重視程度，Bornmann 等人再將 46 篇文獻中有完整量化排序的 38 篇，利用 British Medical Journal 的編輯決策模式進行分析（Howard & Wilkinson, 1998）。表 2-4-2 顯示前 3 名為：結果討論、理論，以及設計與概念，而在表 2-4-1 中領先的貢獻度及寫作與呈現 2 項標準，僅同時名列第 5，而作者聲望或所屬機構則殿後。

doi:10.6342/NTU201602255

表 2-4-2 期刊同儕審查的 9 大評審標準—依文獻中排序分析

標準／理由	標準的重要性			
	高度（1）	中度（2）	低度（3）	平均
1. 結果討論（*n**=30）	50	27	23	1.7
2. 理論（*n**=24）	46	29	25	1.8
3. 設計與概念（*n**=42）	38	26	36	2.0
4. 方法與統計（*n**=37）	27	49	24	2.0
5. 貢獻度相關（*n**=51）	24	37	39	2.2
6. 寫作與呈現（*n**=49）	24	27	49	2.2
7. 文獻分析（*n**=23）	17	40	43	2.3
8. 道德（*n**=12）	8	17	75	2.7
9. 作者聲望或所屬機構（*n**=9）	0	11	89	2.9

* *n* refers to the number of ranking lists in which a main area appears.
X^2（16, *n*=277）= 38.10, $p < 0.001$（based on 10,000 sampling tables）, Cramér's V = 0.26.

資料來源：同表 2-1

Weller（2002）的文獻回顧專書指出期刊編輯是稿件出版與否的主要決策者，並歸納整理出期刊編輯拒絕稿件的 17 項規範標準及 16 項實際標準，發現兩者之間的差異不大，主要有：研究理論與概念、寫作與呈現、研究方法、結果討論、研究分析、研究的重要性等。另外作者亦針對接受稿件進行分析得出 26 項規範標準，發現與拒絕稿件的標準類似，排名前 6 名者為：邏輯嚴謹、貢獻度、研究設計與方法、客觀性、主題適當性、寫作與呈現，至於作者的身份或服務機構均未受到期刊編輯的重視。但是作者亦指出許多拒絕稿件的理由是可以修正的，例如寫作與呈現、資料詮釋或結論、資料分析，以及文獻分析等。

僅有少數文獻利用評審者的審查報告進行研究（Bornmann, Mutz, et al., 2010; Hemlin, 1996; LaFollette, 1992; Shashok, 2008），其中大部份針對單一領域期刊，研究方法與分析重點亦不相同，有檢驗是否存在非科學價值標準（如作者聲望或所屬機構）；有探討評審者的評審風格或寬嚴程度；有評鑑評審作業的效用；還有針對評審者的評語內容進行主題分析（Weller, 2002），發現不同領域期刊的評審重點略有不同，例如化學及麻醉學期刊強調貢獻度及設計與概念（Bornmann, Weymuth, Daniel, 2010; Turcotte, Drolet, & Girard, 2004）；管理學期刊的核心標準為方法論、

doi:10.6342/NTU201602255

分析、理論及貢獻（Robson, Pitt, & West, 2015）等，Chubin 與 Hackett（1990）認為此類研究雖有其通則性（generalization）的限制，但是有助於瞭解同儕審查的內部運作機制。

一般來說，評審者報告中所呈現的訊息是錯綜複雜而且具有多重面向，以評語的正負面向來看，大多數評審者的評語是褒貶參半，Gosden（2003）指出評審者在建議刊登時不見得都是讚美，對建議拒絕刊登者也不一定都是強烈批評。許多研究發現評審者的負面意見超過正面，而且拒絕稿件的負面評語數較接受稿件為多，舉數例如下：Robson 等人（2015）以 *Journal of Advertising Research* 進行研究，發現拒絕或大幅修改稿件主要是負面評語、接受稿件多為正面，小幅修改稿件則是正負參半。加拿大學者 Fagan（1990）利用全國性期刊公開徵求閱讀-語言-素養相關研究的作者提供投稿之審查報告，大多數的樣本來自於加拿大及美國，發現拒絕稿件的負面評語約占 2/3。Bornmann、Weymuth 等人（2010）分析 *Angewandte Chemie International Edition*（AC-IE）拒絕後在其他期刊出版之稿件的評審報告，發現負面評語為正面評語的 6.48 倍（4.99:0.77）。

但是也有研究指出接受稿件的負面評語亦不遑多讓，Bakanic、McPhail 與 Simon（1989, p. 643）以 *American Sociological Review* 的評審報告進行研究，發現拒絕稿件的負面與正面評語的比例約為 4:1，而接受稿件則約為 5:1，作者等人認為「此一情況造成作者理解評審者評語的難度，因為接受稿件與拒絕稿件的評語都同樣具有批判性。」Turcotte 等人（2004）的研究也呼應此一觀點，作者等人在不計寫作相關評語後，*Canadian Journal of Anesthesia* 之接受稿件的負面評語為正面評語的 2.64 倍（6.76:2.56）。

另有少數文獻探討評審字數議題，其一為經濟學領域，Laband（1990）向經濟學期刊文獻作者索取投稿之評審報告，發現評審報告的長度與文獻被引用的情況成正相關，但是作者同時也發現文獻愈長，引用也愈多。其二為廣告學領域，Robson 等人（2015）的研究指出接受稿件的評語平均字數最少，大幅修改的平均

doi:10.6342/NTU201602255

字數最高，兩者之比例為 1:5.08，拒絕稿件與小幅修改稿件的差異不大，約為接受稿件的 3.44 至 3.61 倍。

（二）獎助同儕審查的評審標準

獎助同儕審查的評審過程多採兩階段制，先經內部或外部評審者進行獨立審查後，再由內部評審者與獎助機構重要幹部（或董事會成員）所組成的評審委員會進行會議討論，決定最後獲獎名單（Abdoul, Perrey, Amiel, et al., 2012; Bornmann & Daniel, 2005; Cicchetti, 1991; Ismail et al., 2009; Marsh & Bazeley, 1999; RIN, 2010）。通常獎助機構會事先訂定評審的規範標準，提供內、外部評審者及評審委員參考（Abdoul, Perrey, Amiel, et al., 2012; Geisler, 2000）。

獎助同儕審查的評審標準文獻不多，而且大部分以規範標準為主。近年來各國政府獎助機構的同儕審查作業益加公開透明，才逐漸有較多實際標準的研究，有的根據第一手評審檔案資料進行分析，有的甚至實地觀察評審委員會的討論狀況，茲將有關獎助同儕審查評審標準研究之重要結論略述如後。

1、規範標準研究

Porter（2005）調查訪問經常受邀擔任獎助計畫評審委員的 16 位美國維吉尼亞理工學院教授，歸納出 7 項規範標準：精簡、有組織而且容易讀的寫作方式、符合獎助機構的期望與目標、提出某個重要議題的新觀點、展現申請者的熱誠與承諾、證實申請者對領域瞭解的資訊、具有說服性的初步研究資料，以及可行的工作計畫與合理的預算；此外受訪學者亦強調計畫摘要撰寫的重要性，因為研究計畫的第一印象甚為重要，若是計畫摘要無法吸引評審者，落選的機率甚大。

Abdoul、Perrey 與 Amiel 等人（2012）則是蒐集法、英、美、加、澳、歐盟等共 14 家獎助機構的評審指南，歸納綜整出 9 項規範標準，重視程度依序如下：科學利益、研究方法、可行性、原創性與影響力、財務規劃、創新性、

doi:10.6342/NTU201602255

研究倫理、獎助目標的相關性、寫作品質（見表 2-4-3）。

表 2-4-3 獎助研究機構的規範標準—以 14 家獎助機構分析

規範標準	使用單位 (總數=14)
科學利益（促進公共衛生或科學知識、有利病人及臨床操作、有關科學及醫學內容的文獻回顧等）	14
研究方法	14
可行性（申請者背景、研究計畫、研究環境）	14
原創性與影響力（提供新原始資料或新科技面向、研究結果發表或出版計畫）	13
財務規劃（研究時程及研究經費分配）	12
創新性（新科技、新想法、新理論）	6
研究倫理（病人權利）	6
獎助目標的相關性（計畫書領域及研究者之專長特質符合獎助目標）	3
寫作品質	2

資料來源：“Peer review of grant applications: Criteria used and qualitative study of reviewer practices,” by H. Abdoul, C. Perrey, P. Amiel, F. Tubach, S. Gottot, I. Durand-Zaleski, and C. Alberti, 2012, *PLoS ONE, 7*(9), e46054.

研究者比較前述兩篇研究發現，兩者都重視獎助申請案件的學術創新價值、申請者背景、計畫的可行性，以及與經費的合理性。但是有評審經驗的學者所重視的展現申請者的熱誠與承諾及具有說服性的初步研究資料兩項，並未列於獎助機構的 9 大標準之中。而獎助機構所強調的科學的利益、研究的影響力（研究結果發表或出版計畫等），以及研究倫理等，則非評審學者所重視的審查標準。此外獎助機構列為最後 2 名的獎助目標的相關性與寫作品質，對於評審學者來說卻是獎助計畫審核的重要入門條件。

2、實際標準研究

Abdoul、Perrey 與 Amiel 等人（2012）在法國醫學臨床研究獎助機構（French Academic Hospital Research Grant Agencies, PHRCs）的協助下進行同儕審查作

doi:10.6342/NTU201602255

業的調查及觀察研究，發現該機構的評審者（內部及外部）及決審會議評審委員在審查時大多依據 PHRCs 訂定的 6 大規範標準（科學的利益、原創性、研究方法、可行性、倫理議題及財務規劃）進行審查，不過對於各個標準的重視程度不同。表 2-4-4 顯示內部及外部評審者重視的標準有 4 項：原創性、研究方法、科學的利益，以及可行性；至於決審會議評審委員重視的標準為 3 項：研究方法、原創性，以及與獎助目標的相關性，有關財務規劃只有在研究計畫預算過大時才會提出討論，而可行性的評論則大多針對申請者的背景及過去申請紀錄。Abdoul、Perrey 與 Amiel 等人亦指出內部及外部評審者所重視的 4 項指標均屬個人純主觀性質，可能因此導致較為保守的評審結果或偏好資深研究者的學術馬太效應。此外在決審會議時，通常先由內部評審者逐案報告後才進行討論，報告內容對於評審結果的影響甚大。

表 2-4-4 法國 PHRCs 之內部與外部評審者的實際標準

內部與外部評審者的看法	所有評審者 (*n*=65)	內部評審者 (*n*=38)	外部評審者 (*n*=27)
規範標準的重要性			
每個標準都重要	12	5	7
沒有考量重要性	19	17	2
某些比較重要	34	17	18
最重要的實際標準			
原創性	12	5	7
研究方法	10	5	5
科學的利益	5	3	2
可行性	4	1	3
原創性與科學利益同等重要	1	0	1
研究方法與可行性同等重要	1	1	0
依計畫而定	1	1	0

資料來源：同表 2-4-3

少數學者以評審者的審查報告進行實際標準研究，例如 Langfeldt（2001）指出挪威研究委員會（The Research Council of Norway）的不同領域獎助小組

doi:10.6342/NTU201602255

（醫學與健康組、文化與社會組、科學與技術組及環境與發展組）都訂有不同的規範標準，而且第一階段的外部評審者比較遵守各小組所訂的規範標準，但是決審會議之評審委員則比較自主。Bornmann 與 Daniel（2005）以 1985–2000 年間德國生醫基礎研究獎助機構（Boehringer Ingelheim Fonds, BIF）的博士及博士後研究獎學金申請案件為樣本，發現 BIF 訂定的規範標準有 3 項：申請者的科學成就、研究計畫的原創性，以及未來研究計畫進行的實驗室，但是評審委員最常使用的標準是申請者的科學成就及研究計畫的原創性，至於實驗室的重要性則較少提及。

（三）大學教職聘用／升遷同儕審查的評審標準

大學教職聘用／升遷同儕審查作業的機構性差異頗大，甚至在同一大學之內的不同系所也可能有不同的作法。有的以內部評審為主、有的強調外部審查，有的則是內外部評審兼有。通常內部評審作業採多階段式，可分為科系級、學院級及校級評審會議不等，每個階段的委員會成員有別，也有不同的評審重點。此外無論是內部的評審會議紀錄或是外部評審者的審查意見，通常都被視為機密資訊，不對外公開（Weiser, 2012）。

一直以來研究、教學、服務被視為大學教職聘用／升遷的 3 大評審標準，但是會依據雇用系所的實際需求而有所調整。Miller（1978）的調查指出，美國及加拿大社工碩士班之教職升遷有 5 項評審標準，依序為：教學能力、學術成就、學校服務、社區服務，以及出版品，並且在不同的教職位階還有不同的權重，例如教授及副教授職位則比較重視學術成就。

近年來教職的研究成績益受重視，尤其在卓越期刊出版文獻的能力及獎助計畫的獲獎紀錄，此一趨勢已在學術界引發許多批評與討論。美國柏克萊大學 Center for Studies in Higher Education 主持的 The Future of Scholarly Communication Projetc 指出，大學教職聘用／升遷同儕審查過度重視文獻出版指標，等同將評審工作委由出版社辦理，呼籲學術界應該設計一套更精緻的評審標準，以減少對期刊或大

doi:10.6342/NTU201602255

學出版社的盲目依從，以及對引用量化指標的過度依賴（Harley & Acord, 2011; Harley et al., 2010）。此外 Laudel 與 Glaser（2012）亦發現歐盟研究委員會（European Research Council, ERC）的獲獎者，具有大學教職升遷的優勢。而丹麥研究委員會之獨立研究獎學金的受獎人，獲得專任大學教職的比率為 16%，落選者則只有 9%（Bloch et al., 2014），批評者認為此作法係將大學教職聘用／升遷的評審工作交由獎助機構決定（Harley et al., 2010; van Arensbergen et al., 2014b）。

為了促進學術領域有關個人表現評審的公平性，歐盟第 7 架構支持的跨國合作研究計畫 ACUMEN，整合了同儕審查、書目計量及網路計量 3 種評審方式，提出個人學術表現評審架構－ACUMEN Portfolio，並出版詳細操作綱領，以供獎助計畫或大學教職聘用／升遷等主事機構參考運用。圖 2-4-1 的 ACUMEN 架構，第一部份是受評者的經歷自述（career narrative），第二部分為專業（expertise）、產出（output），以及影響（influence）等 3 項次架構（ACUMEN Consortium, 2014; Tatum &Wouters, 2013）。

doi:10.6342/NTU201602255

圖 2-4-1 ACUMEN Portfolio 個人學術表現評審架構

資料來源：“*ACUMEN Portfolio: Resources for evaluation of individual researchers,*” by C. Tatum and P. Wouters, 2013. Retrieved from http://tatum.cc/wp-content/uploads/Tatum-and-Wouters-ACUMEN@euroCRIS-Porto14Nov2013.pdf

二、同儕審查的評審信度研究

同儕審查是學術界建立共識的機制之一，因此評審信度就甚為重要，Wiley（2008, p. 31）指出：「正如實驗室的研究結果為生物演化過程提供線索一樣，每位評審者的評論背後也存在一個事實，例如評審者因某些理由而不喜歡你的獎助計畫，又如所有評審者都提出相同看法，想必一定是重要且值得信賴的。」Ernst、Saradeth 與 Resch（1993）的實驗研究，將一篇醫學期刊稿件交由 45 位專家進行 8 項科學品質標準評分，在 31 份回收評審報告中，幾乎每一項標準都有極好與極壞

doi:10.6342/NTU201602255

的分數，即使在有關語言（linguistics）標準的意見也不一致。作者等人認為：「利用同儕審查做為評鑑科學文獻的方法並不可靠且可能有偏見，所以該方法本身應該要接受評鑑。」（p. 296）此項研究或許只是一個極端的例子，但是一直以來，同儕審查評審者之間的信度過低，已經引起外界對於審查公平性的質疑（Baxt et al., 1998; Bornmann, 2011a; Cicchetti, 1991; Marsh, Bond, & Jayasinghe, 2007）。不過正如 Williams（1977, p. 131）所述：「評審者意見不一致的理由很多，研究這個問題要站在廣泛的基礎之上。」

（一）同儕審查的評審信度研究

目前同儕審查的信度研究以期刊為主，獎助研究計畫次之，大學教職聘用／升遷則闕如。通常期刊稿件與獎助計畫都是由一群專家進行評審並提出接受或拒絕的建議，如果達到科學標準，至少會有兩位以上的評審者持相同的意見。Cicchetti（1991）的回顧文獻指出，期刊與獎助同儕審查的評審信度有多種統計方式，例如internal consistency、interreferee agreement，或stability across time，最常使用的是單點interreferee agreement，他也定義同儕審查的評審信度（inter-reviewer reliability, IRR）為：「對於同一科學文獻之兩份或兩份以上的獨立評審報告的一致程度。」（p. 120）亦即如果期刊稿件或獎助計畫的獨立評審者之間具有高度一致性時，即具有評審信度。

許多學者利用intraclass correlation coefficient（ICC）來測量評審信度，ICC是一種變異數分解法，其值介於正1與負1之間，其缺點為高信度若加上評審者之間的低變異性，信度可能是基於偶然率，因此有些同儕審查的文獻則是利用Kappa係數來測量評審者之間的審查信度。Kappa係數是有關兩位或兩位以上評審者之間的一致性，如果評審者的意見完全一致時則K=1，如果K值接近0，共識度不會高於偶然率。表2-4-5指出多篇期刊與獎助同儕審查的評審信度文獻，其評審信度值在修正偶然率後，大都介於在0.2與0.4之間（Bornmann, 2011a; Cicchetti, 1991）。另外Bornmann、Mutz等人（2010）以後設分析法研究期刊同儕審查的評審信度，在48

doi:10.6342/NTU201602255

篇文獻中共獲得70筆評審信度之相關係數，ICC／r^2 的平均值為0.34，Cohen's Kappa的平均值為0.17，而根據Fleiss原則，Kappa值在0至0.2之間為信度不高（Fleiss, 1981）。

表 2-4-5 同儕審查的評審信度：期刊稿件及獎助計畫

期刊稿件及獎助計畫	Kappa coefficient/ Intraclass correlation
期刊稿件	
Social Problems (Smigel & Ross, 1970)	.40
Journal of Educational Psychology (Marsh & Ball, 1981)	.34
British Medical Journal (Lock, 1985)	.31
American Sociological Review (Hargens & Herting, 1990)	.28
Physiological Zoology (Hargens & Hearting, 1990)	.28
Journal of Personality and Social Psychology (Scott, 1974)	.26
New England Journal of Medicine (Ingelfinger, 1974)	.26
Law & Society Review (Hargens & Herting, 1990)	.17
Angewandte chemie International Edition (Bornmann & Daniel, 2008)	.15
Angewandte Chemie (Daniel, 1993, 2004)	.14
Physical Therapy (Bohannon, 1986)	.12
獎助計畫	
American Heart Association (Wiener et al., 1977)	.37
National Science Foundation (Solid States Physics) (Cicchetti, 1991)	.32
Heart and Stroke Foundation, Medical Research Council of Canada (Hodgson, 1997)	.29

資料來源：“Scientific peer review,” by L. Bornmann, 2011a, *Annual Review of Information Science and Technology, 45*(1), 197-245. 研究者節錄製表

有些研究指出期刊評審者對於拒絕稿件的一致性較高（Cicchetti, 1985, 1991, 1997; Ingelfinger, 1974），而且為接受稿件的2倍（Weller, 2002）；亦即評審者對於不佳稿件的共識程度較高，對於好品質稿件的認知差異較大（Kupfersmid, 1988）。有些學者認為評審信度受到學科領域的影響，Cicchetti（1991）的回顧文獻指出，

doi:10.6342/NTU201602255

在綜合性或擴散領域，拒絕稿件的一致性較高；而在專門或集中領域，拒絕稿件的共識不高；另有學者認為自然科學及物理學因為有共同的理論概念，故有較高的評審信度，反之在社會科學及人文科學則呈現較低的評審信度（Weller, 2002）；而Eberley與Warner（1990）指出在社會學領域的不同次領域，評審者的信度亦有差異。

不過也有學者持不同看法，Bornmann、Mutz等人（2010）的後設分析文獻亦利用7項共變量（covariates）進行評審信度檢驗，發現只有樣本數及評分系統2項變量與評審信度有顯著相關，其中樣本數較大者以及提供評分系統（rating system）之研究的評審信度較低，作者等人因而質疑小樣本及未提供評分系統研究之可信度，認為可能是小樣本且高信度的文獻比較容易出版，而未說明評分系統者為的研究品質堪慮。作者等人亦指出其他盲審、文獻領域、統計方法、審查目的，以及樣本年限等共變量對於評審信度的影響均不顯著。推論儘管雙盲審查已是許多期刊用以避免評審偏見的作法，但是與評審信度卻未呈現顯著關係，也許因為盲審其實難以做到，因為約有1/4至1/3的稿件，評審者可以知道作者身分，例如自我引用可能透露作者身份，而特定領域的長期研究者也極易辨識。

有關獎助同儕審查的評審信度研究不多，評審者的一致性界於68%至82%之間。其中有2篇是經常被學術界引用的信度實驗研究，其一為模擬實驗，Cole等人（1981）將美國NSF的獲選及落選計畫案各取25件，由美國國家科學院的Committee on Science and Public Policy進行模擬評審發現，有25%的計畫案得到相反結果，作者因此推論有一半獲獎者是隨機產生。其二為真實樣本研究，Hodgson（1997）比較同時向加拿大兩家獎助機構的248件申請案，所得結論與Cole等人（1981）的研究類似，兩家獎助機構的評審一致性為73%，這個百分比雖然是兩個不同機構的合理相關度（r=0.59），但若採Kappa統計則為0.44，共識程度依然不高。另外還有3篇研究，包括：美國NSF的化學、物理及經濟學獎助計畫的整體一致性為68%（Cicchetti, 1991）；德國研究委員會（German Research Council, DFG）的評審者一致性為82%

doi:10.6342/NTU201602255

（Hartmann & Neidhardt, 1990），德國BIF為76%（Bornmann & Daniel, 2005）。

有些研究認為獎助同儕審查適合審查落選計畫，因為評審者對於落選計畫的共識程度較高（Cicchetti, 1991; Wiener et al., 1977），Opthof與Wilde（2009）甚至認為獎助同儕審查無法挑選出最佳或卓越作品，而且整個評審過程容易於受到個人偏見的影響。不過Kostoff（1995）持不同看法，認為同儕審查對於卓越或不佳的研究計畫可以獲得共識，低度共識的情形主要發生在二流作品。

近年來獎助同儕審查的評審委員會的決策過程受到重視，委員會議如何運作以及如何達成共識是學術界甚為關切的議題，有些學者認為面對面的會議討論可能產生群體弱化效應（undesired group effect），而議事規則慣例（如尊重專業或投票表決等）讓評審委員無法堅持個人意見，使得評審結果趨於保守（Luukkonen, 2012; Olbrecht & Bornmann, 2010）；此外這種面對面的會議討論模式也比較容易造成交情偏見（Langfeldt, 2006）。紐西蘭總理科技顧問委員會主席科技顧問P. D. Gluckman（2012）支持這種看法，認為評審會議首先由各委員提出觀點，再經由討論而達成共識，使得評審委員的偏見或否決意見可能成為獎助決定的重要關鍵，而且若在分數落差較大之時，創新想法也不容易獲獎。但是Lamont（2009）持不同看法，認為評審委員會議之審查方式優於採用量化引用數據標準，因為透過不同專家的直接對話、討論而達成的共識，讓科學價值的定義更有彈性；而且評審委員在會議中也可以相互檢驗彼此的態度與行為，使得個人威權比較不會直接影響到審議作業的過程與結果。

此外獎助同儕審查的多階段評審結果之間的一致性與關聯性，也是近年來的研究重點，目前研究結果不一。Bornmann、Mutz等人（2008）利用Latent Markov Model分析德國BIF博士及博士後獎學金的評審檔案，發現第一階段的外部審查對獎助決選的影響最大。但是其他學者的研究卻有不同的結論，有的發現外部審查結果與最後獎助決定之間沒有相關性（van Arensbergen et al., 2014a; van den Besselaar & Leydesdorff, 2009）。

doi:10.6342/NTU201602255

（二）評審信度研究的意義與重要性

理論上同儕審查並非一種隨機過程，所以評審者的判斷會達到某種程度的一致性是可以預期的，亦即有合理的評審信度（Hojat, Gonnella, & Caelleigh, 2003）。不過許多因素可能造成同儕審查的低評審信度，包括評審者的個人立場、領域能力或是評審標準認知等（Eckberg, 1991; Kostoff, 1995），甚至評審態度的嚴格與寬鬆亦會影響評審信度（Bornmann, Mutz, et al., 2010; Eckes, 2004）。而且「如果期刊編輯傾向選不同觀點的評審者，評審者之間的低信度將難以避免」（Weller, 2002, p.197）。Gordon（1977）指出評審信度研究的主要問題之一在於期刊編輯對於評審者有不同的期待，除非編輯提供清楚的評審標準指南，否則評審者的看法可能非常不一致。有些學者認為訂定詳細評審指南有助於提升審查品質以及評審信度（Strayhorn, McDermont, & Tanguay, 1993）；Gottfredson（1978）認為與其強調評審信度，不如針對提升稿件品質的評審信度。

有些學者認為同儕審查的低信度，可能受到評審者個人偏見的影響（Campanario, 1998a, 1998b; Chubin & Hackett, 1990; Cicchetti, 1991; Langfeldt, 2001; Opthof & Wilde, 2009），但是持不同看法的學者認為，高信度只有在單一標準及專斷的評審作業中才可能出現（Bailar, 1991; Harnad, 1985; Hodgson, 1997），而且低信度顯示評審者可以誠實地表達不同的意見，是科學活動中健康且重要的現象（Cole, 1992），此外低信度也反應學術界對於研究前沿的缺乏共識（Cole, 2000; Wood & Wessely, 2003）。Weller（2002）認為評審者信度研究的價值與必要性有其限制，主要是用以改進評審作業的方法，以及編訂評審者指南之用。另外就同儕審查的實務操作面來看，期刊編輯及獎助計畫管理者在選擇評審者時大多採取互補原則，以降低評審者的同質性，獲得廣泛多元的意見，因此評審結果的信度自然會低（Chubin & Hackett, 2003; Rennie, 2003; Stricker, 1991）。

總之同儕審查的低評審信度已是一個事實，重要的是找到的原因，對於評審者的審查報告進行分析是個可行的方式（Bornmann, Mutz, et al., 2010; Fiske & Fogg,

doi:10.6342/NTU201602255

1990; Siegelman, 1991），不過此類研究除了分析個別評審者採用的特殊標準外，也要考量評審者可能有不同的寬嚴評分習慣（Daniel, 1993），而且若根據信度指標來推論任何潛在偏見時，也必需重視研究的效度（Marsh et al., 2008）。Weller（2002）認為評審信度研究大多針對出版建議，並未對評審者的評語進行詳細分析，評審者可能對某一稿件做出相同的評論，但是出版建議卻不相同，因此評審信度研究還要檢驗評審者評語與出版建議的一致性。有學者甚至建議評審者只要提出稿件品質改進的建議，不要做接受或拒絕的出版建議（Armstrong, 1996）。

三、同儕審查的評審公平性研究

學術社群對於同儕審查機制的信賴非常重要，這也是同儕審查正當性的根源所在（Frodeman et al., 2012），但是同儕審查的公平性卻經常受到質疑。一些學者認為評審者也是人類，無法完全排除個人的興趣與野心，並可能為了個人利益而出現偏袒、偏見、忌嫉、疏忽、惡意、腐敗等行為（Abdoul, Perrey, Tubach, et al., 2012; Hames, 2007; Hemlin, 1996; Rennie, 2003; Ziman, 2000）。建構主義學者 Cole（1992）認為學術領域的評審工作是普遍因素及特別因素互動的結果，客觀的評審達成不易，而且影響評審者想法的因素更是難以預期、控制及標準化（Shashok, 2005），因此同儕審查的經常性監督與檢驗，以及深入研究確實有其必要性（Sandström & Hällsten, 2007）。

（一）同儕審查的公平性與偏見

同儕審查的公平性研究是為了提升審查作業的公平性並減少評審偏見（Bornmann, 2011a）也就是所謂的興利除弊。前者的理論基礎來自於 Merton 學派的普遍主義，認為科學品質的評價應該基於客觀的科學標準，不應受到作者或申請者的個人特殊背景因素的影響，只要科學家恪遵普遍主義就會達到評審的公平性（Merton, 1942; Ziman, 2000）。有關減少偏見的研究概念則來自於建構主義學派，該學派質疑有客觀評審的存在，認為科學社會中存在某些因素，誘導或約制科學

doi:10.6342/NTU201602255

家的行為，當評審的決定與稿件內容或申請計畫的科學品質無關時，就會產生偏見（Cole, 1992; Marsh et al., 2008; Sismondo, 1993; Weller, 2002）。

Bornmann（2011a）的回顧文獻指出，學術界已提出超過 25 種同儕審查的偏見，可分為二大類：與研究相關的訊息（如受評者所屬機構的知名度）及與研究無關的訊息（如受評者的性別、年齡或國籍），作者認為前者是造成偏見的主要原因。另有學者將偏見分為好壞兩種，所謂好偏見係指偏好重要的、原創的，以及條理清楚的學術產品，大多數的編輯與評審者都將這種偏見視為理所當然，甚或視為審查的責任。至於導致評審不公的則是壞偏見，主要來自於編輯或評審者對於稿件的來源（作者、機構及國家）、想法與發現有先入為主的個人觀點（Godlee & Dickersin, 2003）。此外 Marsh 等人（2007）認為缺乏評審信度是同儕審查最重要的缺點，有些學者認為稿件不應被評審，因為任何標準均因個人的解讀而有所不同，而一個有價值或可出版的普遍、清晰的潛在範圍是不可能存在的（Bedeian, 2004; Campanario, 1998a; LaFollette, 1992）。

（二）同儕審查的公平性研究現況

同儕審查公平性的研究大多在於探討偏見的潛在來源，主要有兩種研究方式其一為針對同儕審查作業的參與者進行調查；其二為根據評審結果檢視可能的偏見變項（Bornmann, 2011a）。

許多期刊與獎助同儕審查的調查都反應學術社群對評審公平性的質疑。Resnik、Gutierrez-ford 與 Peddada（2008）的調查發現，學者認為期刊同儕審查的評審者不適任者占 68%、有偏見者占 50.5%、要求增加不需要的參考書目者占 22.7%，以及評論含人身攻擊者占 17.7%。英國 Sense About Science（2010）進行的期刊同儕審查全球性調查報告指出，學者對於評審作業滿意及非常滿意者占 69%（2007 年為 65%），儘管不滿意及非常不滿意者僅占 9%，但是支持雙盲評審者高達 76%，顯示學術界對於評審公平性的重視。有關獎助同儕審查的公平性調查，約有 40%的學者不予認同，例如美國 Institute of Handicapped Research 的調查中有 41%的申請

doi:10.6342/NTU201602255

者不認同評審者的評論是公平的（Fuhrer & Grabois, 1985）；美國國家癌症中心（National Cancer Institute）有 40%的申請者認為評審者對於美國某些小型大學或機構存有偏見（Gillespie, Chubin, & Kurzon, 1985）；美國 NSF 有 40%的申請者認為評審品質不完整或不正確（McCullough, 1989）；而澳洲 ARC 的申請者則認為 Large Grant 存在機構偏見及交情偏見，也有利益衝突的問題（Over, 1996）。

第二種研究方法類似利用評審結果進行資料探勘，通常先將受評者（申請者或作者）或評審者的個人特質加以分群，再分析不同屬性的變項與評審結果之間是否存在持續性且系統化的差異。此類研究必需非常審慎且所費不貲，除了需要大量參與者的資料外，也要注意因果推論的合理性（Bornmann, 2011b）。目前相關文獻大多利用評審結果與受評者的屬性進行分析，較少利用評審者特質或評審報告內容進行分析。兩篇著名且被大量引用的公平性文獻，其一為 Zuckerman 與 Merton（1971）的期刊同儕審查研究，發現作者的學術地位有助於稿件的接受度；另一為 Wenneras 與 Wold（1997）的獎助同儕審查研究，發現瑞典醫學研究委員會（Medical Research Council, MFR）的博士後獎學金同儕審查作業，存有女性偏見及交情偏見，此一結論促使 MFR 調整同儕審查評審作業。2007 年，瑞典學者 Sandström 與 Hällsten（2007）再次利用相同方式檢驗 MFR 的同儕審查作業，並未再發現女性偏見，但是交情偏見依然存在，此外還發現生產主義偏見，亦即過度強調預期效用及出版品的影響力。

有些學者試圖將現有之性別偏見文獻通則化，Bornmann 等人（2007）利用後設分析法發現獎助同儕審查對女性存有偏見。但是 Marsh 等人（2008）分析澳洲 ARC 一萬餘筆獎助申請的評審檔案，發現在不同領域的獎助計畫中均未存在性別偏見。為了進一步瞭解性別偏見的真相，前述兩篇文獻的研究團隊合作以不同的後設分析方式進行研究，所據文獻包括 8 國超過 35 萬筆獎助申請資料，發現獎助計畫同儕審查作業，無論是自然或人文科學領域，都沒有存在性別偏見；但是在獎學金同儕審查上，則對男性申請者較為有利，不過因為個別文獻的差異性甚大，

doi:10.6342/NTU201602255

仍需更進一步分析（Marsh & Bornmann, 2009）。Marsh 等人（2011）認為偏見的通則性推論必需非常審慎，二手資料的後設分析及大規範的一手資料研究各有優點與限制，建議兩種同時使用才可以獲得較為可信的結論。

總之，同儕審查的公平性研究囿於資料取得不易，而且因果推論困難而進展有限，不過長期、定時的監督與檢驗可能是唯一解決之道，瑞典醫學研究委員會（MFR）調整評審作業而改進對女性的偏見，就是一個最佳的例子，此外，同儕審查的評審過程若能更加公開與透明，應該可以有助於提升評審作業的公平性。Wenneras 與 Wold（1997）認為學術評鑑系統必需接受科學的檢驗，以免引起外界的質疑，而如何建立一個可以排除個人偏見且公平的同儕審查系統，乃是學術界的當務之急。不過也有學者認為「同儕審查若不可避免受到社會性及不公平性的影響，是否仍宜標舉審查之公平性。」（Lee et al., 2013, p. 13）。

第五節　同儕審查品質提升的多元作法—以期刊為例

半個世紀以來，期刊同儕審查的研究主要集中在事後的檢驗與監督，重要議題包括公平性、信度，以及效度等，不過如何提升評審品質也是期刊出版者與學術界甚為重視的議題，相關研究與討論亦眾，有的分析優良評審特質，有建議提供評審者訓練，更有建置評審者評鑑系統之作法，不過如 Callaham（2013）所述「優良評審者如何形成，仍然未有定論。」

一、優良評審者的特質與評審訓練

對於期刊編輯來說，評審者的主要功能有二，其一為協助編輯決定接受或拒絕稿件，其二為協助改進稿件的品質（van Rooyen et al., 1999），高品質的評審是期刊同儕審查的成功要件，更是期刊編輯追求的目標。每一位期刊編輯都希望在選擇評審者之時能有更多的參考訊息，以確保評審品質，並避免參考價值有限的

doi:10.6342/NTU201602255

劣質評審。

有些研究探討評審者的特質與評審表現的關聯性，包括評審者的學術地位、職業位階、年齡、服務機構，以及評審經驗等（Evans, McNutt, Fletcher, & Fletcher, 1993; Friedman, 1995; Murphy & Utts, 1994; Nylenna et al., 1994; Stossel, 1985），不過各篇文獻的研究結果不一，有些文獻認為有經驗的評審者表現較佳，有些則認為年輕且位階較低的學者是最佳評審者（Callaham, 2013; Weller, 2002）。

另外有些學者認為評審的能力不是天生而來，應該給予適當的訓練。目前大多數的評審者都未經正規訓練，少數指導教授以一對一的方式帶領學生瞭解同儕審查的過程（Caelleigh, Shea, & Penn, 2001），大部份的評審者則是邊做邊學（Tsang & Frey, 2007）。Garrow、Butterfield、Marshall 與 Williamson（1998）亦指出甚至期刊編輯也缺乏正式訓練，他們大多是師徒制，由年輕編輯向資深編輯學習。最近的調查亦指出，主要臨床醫學期刊的編輯對於同儕審查的知識不足（Wong & Callaham, 2012）。

不過已有研究指出短期的同儕審查訓練課程或研習會，對評審品質的提升效用有限，推論可能是訓練課程太短，並建議進一步檢驗長期課程的效用（Callaham & Schriger, 2002; Callaham, Wears, & Waeckerle 1998; Schroter et al., 2004）。De Vries 等人（2009）認為相對比較簡易的學習方法是將同篇稿件的其他評審者的審查報告以及期刊編輯的決定信函提供每位評審者參考，評審者可以彼此學習。

儘管如此，英國人文社會科學院的報告仍然強調訓練課程對於年輕研究者的重要性，認為同儕審查係依賴專業規範執行，因此需要有適當的訓練課程予以強化，以確保評審者的能力；報告中亦建議同儕審查的訓練要納入高等教育的課程之中，訓練的內容不只強調學術品質對於科學發展的重要性，也要認知在同儕審查作業中的專業道德問題，包括尊重著作權以及公平的對待他人的創作等議題（BA, 2007）。

doi:10.6342/NTU201602255

二、評審的課責性

理論上大多數的期刊編輯在將稿件付交評審時，都會提供評審指南或制式的審查表格，根據研究大多數（89%）的評審者偏好結構式表格，不贊成簡短或參考價值有限的評審指南，而沒有經驗的評審對於結構式表格的偏好更高達 97%（Nylenna et al., 1994; Nylenna, Riis, & Karlson, 1995）。

有些期刊編輯及學者認為評審者是一個同時具有特權與責任的工作，除了專業知識外，評審者做為服務期刊與作者的角色，應具有負責任的態度與適當的禮節，Drotar（2009）認為評審報告要具有清楚性、專指性、建設性及完整性，有些學者提出優良評審報告的特色（Drotar, 2009; Lovejoy, Revenson, & France, 2011），研究者綜整五大特色如下：

（一）合宜的評審態度：評審者要維持專業及尊敬的語調，貶抑的語言將使評審目的失焦，不過也要勇於指出稿件的缺點。

（二）保密的出版建議：不可告知作者出版建議，以免增加編輯的困擾。

（三）評語內容的合理性：

1、評語的一致性：除了致作者及致編輯的評語要一致外，評語的內容也要與出版建議配合，例如評語若是對稿件多方讚美，但是卻做出大幅修改或拒絕稿件之建議，對編輯來說將難以適從。

2、區分主要評語與次要評語：主要評語是稿件拒絕或接受的重要理由，就拒絕稿件來看應是「無法修改」的缺點，對於接受稿件則是出版前「必需修改」的問題；至於次要評語則是有關稿件的可讀性或清楚性，不過評審者也應理解，次要評語主要是期刊編輯的工作，而不是評審的焦點。

3、審慎提出修改建議：如果稿件內容經過改正可以出版，就需提供詳細的意見，如果是建議拒絕之稿件，只要描述最基本的錯誤及不可改正的缺點即可。此外評審者也不應強迫作者增加評審者認為有興趣的內容，但是卻不屬於稿件研究的主要議題。

doi:10.6342/NTU201602255

（四）適當的請求協助：評審者並非「絕對」需要評論稿件的每一部份，尤其是當評審者感到某些章節如技術、方法或統計分析等，已經超出自身的專業，但是必須明白告知期刊編輯另尋其他評審者協助。

（五）易讀的評語格式：包括開場段落描述評審者對稿件的整體意見，條列式意見以利作者回復修改情形，以及明確指出評語所在的確切文字，儘量不要使用模糊或概括式論述。

此外 Bormann、Weymuth 等人（2010）以德國化學期刊 *Angewandte Chemie International Edition*（AE-IE）的拒絕後在其他期刊刊登的稿件進行研究，發現在評審報告中若是存在貢獻度相關及設計與概念之負面評語，稿件則無法被其他影響係數較高的期刊接受，因而建議化學領域期刊稿件之負面評語若無涉貢獻度相關或設計與概念，評審者應建議修改而不宜直接拒絕，而期刊編輯也應考量讓作者修改或重新審查。本研究認為此種自省式之作法，應有助於強化期刊同儕審查的課責性。

三、評審者的評鑑

科學出版品依賴同儕審查作業，希望評審者以周延、有效，以及一致性的態度審查稿件。這個作業最主要的缺點在於評審者的多變性以及差異懸殊的評審品質（Burnham, 1990; Ingelfinger, 1974）。Baxt 等人（1998）認為評審者多變性的理由之一也許是因為大多數的期刊對於評審者的選擇、訓練，以及評鑑缺乏有組織且正式的過程，許多期刊選擇評審者的方式是非正式或前後不一致的，而且大都基於過去的接觸經驗。2007 年一項針對美國 3000 位科學家進行的調查指出，80% 的評鑑工作由少數專家進行，也就是少數的學者負責大量的評鑑工作，調查者建議必須進行同儕審查的評鑑以改進其效用，亦是永續經營的保證（Alberts, Hanson, & Kelner, 2008）。

雖然某些著名期刊對於擔任評審者的門檻甚高，所以比較可以確保審查品質，

doi:10.6342/NTU201602255

但是目前全球大致有 60%的期刊評審者是義務性質，大多數期刊面臨的狀況經常是評審者難尋，而且也不檢驗評審者的表現（Good, Parente, Rennie, & Fletcher, 1999）。許多期刊編輯自認相當瞭解評審者的優點與表現，但是有些學者認為這種情況只有在相對小規模的期刊才可能出現，對於擁有大量評審群或綜合性期刊，完善的評審者評鑑系統確實有其必要性。

已有研究證實由編輯主觀評分的評審者評鑑系統的效用，發現評鑑結果與評審者審查稿件錯誤的能力呈正相關（Callaham, Baxt, Waeckerle, & Wears, 1998; Feurer, Becker, Picus, Ramirez, Darcy, & Hicks, 1994; van Rooyen et al., 1999）。Callaham（2003）認為擁有超過 100 位評審者的期刊，就需要建置一個由編輯所維護的評審品質評鑑系統，以追蹤評審者的表現，但是評鑑系統的設計不能太複雜，若要求編輯付出太多額外的時間與精力，評鑑的效用即可能打折。期刊編輯可以利用這個系統定期（通常一年一次）檢視相關報表，對於持續低分的評審者可以將之除名或要求改正；最好的評審者則可以給予獎勵，以及聘請協助訓練或指導其他的評審者。

此外也有學者認為以不同稿件來評鑑評審者的表現，有其比較上的限制，因為每篇稿件最多只有 3 至 4 位評審者，應該讓所有的評審者針對同一稿件進行審查，才能系統比較評審者的表現。Baxt 等人（1998）在一篇假造稿件中設計了 10 個大錯誤及 13 個小錯誤，寄交 Annals of Emergency Medicine 的 262 位評審者進行盲測，在 78%擲還之評審報告中發現，評審者除了對出版建議的看法不同外，所指出的錯誤亦大相逕庭，而且有 2/3 的主要錯誤沒有被指出，有 68%的評審者未能發現結果討論並不支持研究結論。該期刊每年接受 800 件投稿，稿件的接受率為 26%。

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

第三章　研究設計與實施

本研究蒐集 3 種我國科技部（前身為國家科學委員會）評定之 A 級或 B 級的社會暨人文科學期刊[1]、1 年或 2 年之同儕審查的評審者審查報告，利用內容分析法探索評審者使用的實際標準，並檢測評審者彼此之間的信度，以及討論評審的公平性與課責性議題。本章的第一節概述研究方法；第二節為研究對象與資料蒐集；第三節為研究設計，包括研究單元、內容分析系統、編碼原則與結果舉例，以及評審信度統計；第四節說明內容分析法之信度與效度；第五節為研究步驟，茲分別說明如後。

第一節　研究方法

本研究採用主題分析（thematic analysis）式內容分析法，將期刊同儕審查之評審報告內容依不同主題加以分類，此一方式在內容分析研究甚為常見，不過運用在評審者評語的文獻不多（Bornmann, Herich, Joos, & Daniel, 2012）。

所謂內容分析是一種蒐集與分析文本內容的技術，將文本中分散的定性符號轉變為具體的定量資料，做為統計分析的依據。內容分析法在操作上屬於非反應式的量化研究技術，被研究的人並不知道他們是某個研究計畫的一部分，研究者也不可能影響到原始創作者與接收者的溝通過程（Neuman, 2003）

Berelson（1952）強調內容分析法必需具有客觀性、系統性與定量性。所謂客觀性是指進行研究時必須有一套含意明確且界限條理清楚的規則引導，以降低個人的主觀立場；系統性是指內容和類目的採用或捨棄，必須依據始終一致的法則；定量性則是依規則對擬定之類目與分析單位加以計量。20 世紀末期，因著電腦科

[1]中華民國科技部（前身為國家科學委員會）在 2014 年完成社會暨人文科學期刊評比，列名 A 級者有 111 種、B 級者有 96 種（合計 207 種，約占期刊總數 20%），本研究的 3 種期刊分屬於 A 級或 B 級期刊。

doi:10.6342/NTU201602255

技的進步及統計理論的發展，除了拓展內容分析法的應用層面，並強化內容分析的解析能力與準確性（楊開煌，1998）。

總而言之，內容分析法讓研究者透過編碼系統將代表變項的內容層面轉換成數字之後，就開始採用與實驗法或調查法相同的方式進行統計分析 （Neuman, 2003）。Rubin 與 Babbie（2011）認為內容分析法雖然在推論與解釋上具有質性研究的成分，但是其基礎發展仍然是以量化概念出發，可算是一種質量並重的研究方法。

第二節　研究對象與資料蒐集

本研究採便利抽樣，經聯繫多家我國政府評定為 A 級或 B 級之社會暨人文科學期刊，在說明本研究之目的及資料使用方式後，共有 3 家綜合或專門期刊願意提供評審者的審查報告，樣本資料來源如下：1 家 A 級期刊（拒稿率大於 50%）提供 2012 年的審查報告、2 家 B 級期刊（拒稿率低於 50%）分別提供 1 年（2004 年）及 2 年（2006 年至 2007 年）的審查報告，至於由期刊編輯直接拒絕的稿件並未包括在內。

本研究所蒐集之期刊稿件共 48 篇，各篇之評審者由 2 位至 3 位不等，合計有 103 份評審報告。各期刊所提供之評審報告均已先經過拆解，刪除評審者姓名、稿件作者或標題等涉及個人背景的訊息，亦不包括期刊編輯的最後出版決定，這個拆解主要是為了機密性，但是仍然可以呈現評審者的評語與出版建議之間的關係。

doi:10.6342/NTU201602255

第三節　研究設計

一、研究單元

本研究之研究單元為一審的評審報告內容，針對評審者致作者之評語及致編輯之出版建議進行分析。使用一審報告之原因有二，其一、修改或重投稿件的評審內容大多仍然限於一審意見範圍；其二、有些樣本亦包括評審者致編輯之評語，惟內容多與致作者評語雷同，只是文字較為簡潔，故不納入分析。另本研究利用稿件之研究類型（量化研究及非量化研究）及稿件領域（社會科學及人文科學）進行交叉分析。

二、內容分析系統

本研究之內容分析系統計分為三大項，包括評語主題分類架構與評分系統、潛在公平性類目，以及交叉分析變項，詳述如後。

（一）評語主題分類架構及評分系統

為利與相關文獻進行分析比較，本研究之評語主題編碼是採自 Bornmann、Nast 等人（2008）的後設分析文獻所提出之期刊同儕審查評審標準分類架構[2]。該架構共分為 9 大類、35 小類，大類主題與內容略述如後，表 3-1-1 為本研究之評審報告編碼舉例。

1、貢獻度相關（Relevance of contribution）：稿件對於科學發展與期刊讀者的價值，或是在實務上的貢獻，強調稿件的重要性、創新性，以及原創性，此外研究結論相關評語亦包括在此類。

2、寫作與呈現（Writing/Presentation）：與稿件形式品質相關之評語，例如寫作風格與呈現、拼字、文法，以及專業性；稿件的完整性，例如稿件在不同的章節之必要資訊，完整清晰的呈現，此外稿件是否符合期刊投稿

[2] Bornmann 為該文獻的通訊作者，研究者在聯繫後獲供原始分類檔案（見附錄一）。

doi:10.6342/NTU201602255

指南或是長度是否適當亦屬此類。

3、設計與概念（Design/Conception）：正確且邏輯性的概念架構及適當的研究設計，包括研究內容的一致性，配合研究問題的研究設計、樣本的品質、結果的通則性，以及可複製性。

4、方法與統計（Method/Statistics）：研究方法或統計分析的正確性、適當性，以及新穎性；此外研究方法之操作與測量品質亦包括在內。

5、結果討論（Discussion of results）：基於分析結果之客觀、正確，以及適切的討論、清晰易懂的內容，以及詮釋的深度。

6、文獻分析（Reference to the literature and documentation）：與相關研究的關係，包括文獻的涵蓋面，以及文獻分析的正確性。

7、理論（Theory）；具有對理論發展的貢獻，包括新創理論、舊理論新解，此外缺乏理論詮釋，或不正確使用概念或理論架構等亦屬此類；本研究未發現此類主題之評語。

8、作者聲望或所屬機構（Author's reputation/Institutional affiliation）：前 7 大類是用來評審稿件的內容，本類強調作者在其領域的聲望，以及所屬機構；本研究未發現此類主題之評語。

9、道德（Ethics）：稿件符合領域之科學道德或道德標準，例如重複出版、二手分析，以及領域道德；本研究未發現此類主題之評語。

doi:10.6342/NTU201602255

表 3-2-1 評語主題編碼舉例

評語分類	評語舉例
貢獻度相關	「題材具有時事性…」;「研究成果非常具有實務性且貢獻良多…」;「主題的選擇有意義…」;「提出了與以前研究的不同說法，值得學界參考…」;「本文目前內容相當一般，未能突破前人研究…」;「…沒有提出明確的觀點與貢獻…」;「結論略顯空泛…」;「有些結論與研究目的不一致…」;「看起來像是一篇…教學報告」;「所言皆屬常識，且幾乎都是…先行研究的拼湊」;「根據二、三手材料…作以偏概全的價值論斷」。
寫作與呈現	「論文的結構清楚，說明也很詳盡…」;「結構嚴整、所使用的文字精準流暢、說理精確…」;「比對十分細密、行文堪稱流暢…」;「中英文摘要語意不清楚…」;「用辭遣字可再斟酌…文字宜更為謹慎」;「標點符號與格式有錯誤…應重新檢查」;「英文摘要之文法錯誤相當多…」;「可能有一些錯字，請再校對一遍…」;「章節格式稍感複雜，閱讀後不易抓到作者欲陳述的重點…」;「資料紛雜，行文有如流水帳，不易理清頭緒…」。
設計與概念	「問題提出十分清楚…分析對象是研究典範轉移的最佳題材」;「具有明確的方法論之自覺…全文問題意識明確…」;「本文選用的討論文本頗為豐富…」;「增加樣本的多元性…使研究結果與貢獻更臻完善」;「主題茲事體大，很難以一篇單篇論文論述清楚…」;「本文應從整個大體來梳理…」;「沒有明顯精確的研究問題…」。
方法與統計	「研究方法、步驟尚稱嚴謹…」;「以評量表統計分析方式，進行實質內容屬性客觀的歸納…」;「運用一個信度良好的研究工具進行實徵研究，並能確實執行…」;「問卷內容...未依研究對象做修正...」;「問卷設計有幾個題目設計不太適合…」;「就資料選樣來說，僅以短時間的取樣應不足推論母體…」「質性訪談之問題部分偏離主題…」;「量化研究過程及量化結果與比較分析過於簡約…」。
結果與討論	「綜合討論非常深入，值得肯定…」;「舉例佐證，有本有源…」;「概念分析、原典的引用與解說，都十分清楚正確…」;「針對研究研究方法的疏漏進行探討並做出解釋…」;「建議文章能補充更深入的分析及討論…」;「若能再加強質化分析詮釋，會更具參考價值…」;「研究結果只是報告出來…」;「提供十分豐富的例證支持論文的論點…」。

doi:10.6342/NTU201602255

評語分類	評語舉例
文獻分析	「文獻探討完整…」;「相關文獻…析論深入」;「引用方式、引用文獻品質等仍有些不足之處…」;「文獻探討與主題規劃不一致…」;「…作者應蒐集文獻做深入的討論」;「作者引用文獻資料極為貧乏…」;「已有不少研究成果，作者全未引述說明…」。
理論	（無相關評語）
作者聲望或所屬機構	（無相關評語）
道德	（無相關評語）

另為探討評審者的評語與其出版建議之一致性，研究者自創評語評分系統，針對正面評語之「具體性」(評語內容明確且專指）及負面評語之「可修改性」(評語內容之修改難易程度）進行正負 3 級距分析。「具體性」為：不具體（1)、尚可（2)，以及很具體（3);「可修改性」為：困難修改（-3)、尚可（-2)，以及容易修改（-1)。

（二）潛在公平性類目

本研究樣本資料未包括評審者及作者背景等資料，僅針對評審報告內容來探討公平性議題，研究者參考利用評審報告進行之公平性研究文獻（Beyer, Chanove, & Fox, 1995; Caligiuri & Thomas, 2013; Gilliland & Cortina, 1997; Lovejoy et al., 2011; Spencer, Hartnett, & Mahoney, 1986）以及某些期刊編輯或學者提出之期刊同儕審查指南（Caligiuri & Thomas, 2013; Lovejoy et al., 2011)，建置 3 項潛在公平性類目，包括非科學品質評語、情緒性用語，以及向作者透露出版建議。其中向作者透露出版建議一項，係因我國為小型科研國家，各領域學門的研究者不多，評審者比較容易辨識出作者身份，而向作者透露出版建議可能存在評審公平性的問題。

有關非科學品質評語的操作型定義，係指除了科學品質之外的評語，例如作者姓別或個人所屬機構等；至於情緒性用語的操作型定義，主要是指評審者的評

doi:10.6342/NTU201602255

語帶有貶抑態度，有違專業與尊敬的溝通語調，例如「研究問題與概念風馬牛不相及...」;「內容討論實在有點不知所云...」;「戴著有色眼鏡分析...」;「所犯錯誤離譜者不在話下...」;「解讀文獻資料有牛頭不對馬嘴的驚人結論...」;「研究無新意換湯不換藥...」等。

（三）交叉分析變項

1、出版建議：分為拒絕稿件、大幅修改稿件（再審或接受）、小幅修改稿件（接受），以及接受稿件等 4 類。

2、稿件研究類型：區分為量化研究及非量化研究兩類。根據醫學領域學者 Hojat 等人（2003）認為期刊稿件審查應該分為實徵研究（量化研究）、非實徵研究（如質性研究、理論或概念分析，以及文獻回顧等），以及雜文（投書及評論等）3 類，其中實徵研究因有相同的評審標準（如研究設計、樣本、研究方法及統計分析等），評審原則比較具有一致性。另外心理學者也強調實徵研究、理論討論，以及文獻分析等稿件各有不同的審查重點與技巧（Sternberg, 2005）。鑒此，本研究順應國內學術界慣例，分為量化研究及非量化研究進行討論。

3、稿件領域：分為社會科學與人文科學 2 類。本研究蒐集之 3 種期刊包括綜合性及專門期刊，因此樣本稿件所屬學門廣泛，單一學門的樣本數不足，故以領域進行分析。

三、評語編碼原則與結果舉例

本研究之編碼分析單位至少是完整的句子（Gosden, 2003），若涉及多項評審標準時可重複分類（Bornmann, Nast, et al. 2008），評語之主題分類系統係依據 Bornmann、Nast 等人（2008）提出之期刊同儕審查評審標準分類架構。每項評語除了分類外，亦將正面評語根據「具體性」（評語內容明確並專指）、負面評語根據「可修改性」（評語內容之修改難易程度）進行正負面向之三級距評分。此外對

doi:10.6342/NTU201602255

於評審者使用非科學品質評語、情緒性用語，或者向作者透露出版建議等三項潛在公平性類目也加以編碼，以利分析公平性議題。

表 3-2-2 節錄自編碼表單，以貢獻度相關及結果討論兩項標準為例，表格之橫列為每位評審者所提出之評語編碼情形，表格縱列為評審者評語所屬主題類別及正負評分。

表 3-2-2 評語分類與評分之編碼舉例—以貢獻度相關及結果討論為例

稿件編號	評審者編號	評語之分類及正負評分				
		貢獻度相關			結果討論	
		評語一	評語二	評語三	評語一	評語二
1	1	1	-2	1	1	-2
	2	2	-1	1	1	-1
2	1	1	-1	-3	-2	1
	2	2	1	-1	2	-1
	3	-1	-3	1	-1	-2

備註：1、正面評語以「具體性」分為 3 級：不具體（1）、尚可（2），以及很具體（3）；2、負面評語以「可修改性」分為 3 級：困難修改（-3）、尚可（-2），以及容易修改（-1）。

根據其他使用同一分類架構的文獻發現，大類評審標準下之小類會出現正負評語夾雜的情況（Bornmann et al., 2012; Bornmann et al., 2010），本研究亦然；究其原因為大類下之小類各有其獨特屬性，評審者可能對同一大類之不同小類提出正負不同的意見，以貢獻度相關為例，其小類包括稿件主題的重要性，科學貢獻、創新性、實務價值、符合期刊領域屬性，以及結果貢獻度等，當評審者在肯定稿件之創新性或重要性時，也可能同時批評結果與建議之論述不足，而造成同一大類下的小類評語有褒有貶。

四、評審信度統計

Cicchetti（1991）的回顧文獻將同儕審查的評審信度（IRR）定義為：「對於同一科學文獻之兩份或兩份以上的獨立評審報告的一致程度（p. 120）。」也就是如果期刊稿件的獨立評審者之間具有高度一致性時，即具有評審信度。許多學者利

doi:10.6342/NTU201602255

用不同的統計方式研究評審信度，目前大多採用ICC或Cohen's Kappa係數（Bornmann, 2011a）。

本研究利用Cohen's Kappa係數針對評審者的出版建議及所使用之評審標準進行信度分析，Cohen's Kappa係數是關於兩位或兩位以上評審者之間的一致性，其公式如下：

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

P_0：實際一致性（actual agreement），前後（兩種）測量結果一致的百分比

P_c：期望一致性（chance agreement），前後（兩種）測量結果預期相同的機率

K 值計算的結果為-1 至 1，但通常 K 值落在 0 至 1 之間，可分為五組來表示不同等級的一致性：0.00 至 0.20 為極低的一致性（slight）、0.21 至 0.40 一般的一致性（fair）、0.41 至 0.60 中等的一致性（moderate）、0.61 至 0.80 高度的一致性（substantial），以及 0.81 至 1 幾乎完全一致（almost perfect）（Fleiss, 1981）。舉例來看，Daniel（1993, p. 23）指出：「若以 Kappa 係數 0.23 為例，表示評審之間對於 23%的文稿內容的評價趨於一致，而不只是隨機的結果。」

第四節　內容分析法之信度與效度

本研究的評審報告內容編碼工作，係由研究者及一位國立台灣大學圖書資訊所的博士生共同合作。在進行分類編碼前，兩位編碼員先進行密集會商，針對各項分析類目之定義與歸類原則等進行討論，例如在人文科學的非量化研究中，許多是針對某一概念進行論述（或辯證），本研究將評審者對於整體概念的評語歸類於「設計與概念」，而對於論述內容之意見則歸於「結果討論」；另外在許多社會科學的實徵研究中，有些評審者雖然肯定研究之重要性或科學價值，但是會針對

doi:10.6342/NTU201602255

研究之「結果與建議」提出補強之建議，本研究則將之歸於「貢獻度相關」大類之下的小類「結果貢獻度」。

本研究的分析類目大多採用目前研究之結論，故不再進行效度測量。此外本研究之編碼流程是由兩位編碼員各自編碼，再彼此檢驗編碼結果；因已在事前針對類目之定義與編碼原則進行充分溝通，編碼之一致性達 86%，至於差異之處則經過充分討論後取得共識。

第五節　研究步驟

本研究之研究步驟如下：

一、文獻蒐集與分析

透過相關文獻之蒐集、閱讀與分析，首先回顧同儕審查機制的起源、定義與優缺點，以及檢驗與發展，以探討同儕審查做為知識研究主體的歷程。其次分析期刊、獎助、大學教職聘用／升遷同儕審查之評審標準、信度與公平性之研究現況，並彙整各界為提升期刊同儕審查之評審品質的多元作法，以利進行我國期刊同儕審查之研究。

二、擬定研究目的與範圍

閱讀相關文獻釐清同儕審查之研究現況與重要議題，並確定本研究之研究目的與範圍。

三、規劃研究設計

參考文獻分析之結果，選擇適合之研究方法，並著手規劃研究設計。本研究之實徵研究係以我國社會暨人文科學之期刊同儕審查為例，針對評審者的評審報告，以內容分析法探討評審者在審查時使用之實際標準，並將進行評審者之間的

doi:10.6342/NTU201602255

信度分析，以及討論評審的公平性及課責性議題。

四、資料蒐集

本研究除了回顧同儕審查之起源、研究現況與發展外，並以期刊同儕審查之實徵研究為例，利用 3 種我國科技部（前身為國家科學委員會）評定之 A 級或 B 級的社會暨人文科學期刊之同儕審查評審報告進行研究。有關評審者及作者之個人背景、稿件的主題，以及編輯的出版決定等，均不在資料蒐集範圍之內。此外為保護資料來源，本研究亦不針對個別期刊的特色進行分析與檢驗。

五、資料整理與分析

參考目前研究文獻建置期刊同儕審查評審報告之內容分析系統，由研究者及一位國立台灣大學圖書資訊所博士生合作進行編碼,兩位編碼員首先獨自編碼後，再彼此檢驗一致性，對於不同的編碼結果則以討論方式達成共識。

六、探討評審標準、信度與公平性

根據內容分析編碼結果，瞭解期刊同儕審查的評審者在審查稿件時所使用之實際標準，並討論評審者之間的信度與公平性議題，各議題均以出版建議、稿件研究類型，以及稿件領域進行交叉分析，亦與目前研究結果進行比較，以呈現我國社會暨人文科學期刊之同儕審查的評審特色。

七、探討同儕審查的課責性

鑒於評審報告乃期刊編輯決定稿件出版與否的重要參考資訊，為探討評審報告對期刊編輯之參考價值，本研究對於正面評語之「具體性」以及負面評語之「可修改性」進行量化分級，以探討評審評語與出版建議之一致性，並據以討論評審的課責性議題。

doi:10.6342/NTU201602255

八、撰寫研究結果與建議

根據前述研究程序與步驟，逐步執行資料蒐集與分析工作，彙整並依據所得之研究數據，轉換為具體研究成果，以達成本研究之研究目的，並在最後提出研究結論，以及針對我國同儕審查的整體發展方向，以及我國社會暨人文科學期刊同儕審查的未來研究提出建議。

doi:10.6342/NTU201602255

第四章　研究結果

本研究採便利抽樣，共蒐得 3 種我國科技部（前身為國家科學委員會）評定之 A 級或 B 級的社會暨人文科學期刊[3]、1 年或 2 年之同儕審查的稿件共 48 篇（不含由期刊編輯直接退稿者）；各篇稿件的評審者有 2 位至 3 位不等，合計 103 份評審報告。本研究利用內容分析法將評語分類及給予正負評分，並以出版建議、稿件研究類型，以及稿件領域進行交叉討論，探索評審者在審查過程中所使用的實際標準，並檢測評審者彼此之間的信度，以及討論評審的公平性與課責性議題。本章研究結果分為樣本概述、評語筆數與字數、評語正負面向及評審標準、評審信度、評審的公平性及課責性等節進行分析，敘述如後。

第一節　樣本概述

本研究以社會暨人文科學的期刊為研究對象，共有 48 篇稿件，其中社會科學有 19 篇，人文科學有 29 篇。稿件中有為數不少的非量化研究，表 4-1-1 顯示在社會科學領域有將近 1/3（31.58%）的非量化研究，而在人文科學領域有 82.76%。有些學者認為期刊稿件審查應該區分為實徵研究與非實徵研究（如非量化研究、理論或概念分析，以及文獻回顧等），其中實徵研究的評審標準比較具有一致性，而非實徵研究亦有不同的評審重點與技巧（Hojat et al., 2003; Sternberg, 2005），因此本研究亦將稿件研究類型納入分析變項討論，惟依國內學術界慣例，區分為量化研究與非量化研究兩類。

[3] 中華民國科技部（前身為國家科學委員會）在 2014 年完成社會暨人文科學期刊評比，列名 A 級者有 111 種、B 級者有 96 種（合計 207 種，約占期刊總數 20%），本研究的 3 種期刊分屬於 A 級或 B 級期刊。

doi:10.6342/NTU201602255

表 4-1-1 樣本稿件之領域及研究類型分析（*n*=48）

稿件領域	稿件研究類型					
	量化研究		非量化研究		總計	
	篇數	%	篇數	%	篇數	%
社會科學	13	68.42	6	31.58	19	100.00
人文科學	6	17.24	23	82.76	29	100.00
總計	19	39.58	29	60.42	48	100.00

本研究 48 篇稿件樣本的評審者為 2 位至 3 位，評審報告計有 103 份，表 4-1-2 顯示評審報告之領域及研究類型，其分配與樣本稿件相差不大，不過在社會科學及人文科學中的量化研究的比例均有提升，推論量化研究稿件之評審者的出版建議較不一致，因此有第 3 位評審者。

表 4-1-2 評審報告之領域及研究類型分析（*n*=103）

稿件領域	稿件研究類型					
	量化研究		非量化研究		總計	
	份數	%	份數	%	份數	%
社會科學	31	72.09	12	27.91	43	100.00
人文科學	12	20.00	48	80.00	60	100.00
總計	43	41.75	60	58.25	103	100.00

本研究的 103 份評審報告經過內容分析後得到之評語數共 466 筆（涉及多項標準者可重複分類），表 4-1-3 為評語之領域及研究類型分析，與表 4-1-2 相較，社會科學之量化研究及人文科學之非量化研究的評語筆數百分比均略有增加。

表 4-1-3 評語之領域及研究類型分析（*n*=466）

稿件領域	稿件研究類型					
	量化研究		非量化研究		總計	
	筆數	%	筆數	%	筆數	%
社會科學	166	75.11	55	24.89	221	100.00
人文科學	44	17.96	201	82.04	245	100.00
總計	210	45.06	256	54.94	466	100.00

doi:10.6342/NTU201602255

而以 103 份評審報告之出版建議分析，表 4-1-4 顯示拒絕稿件約占 1/5(19.42%)，大幅修改稿件占 29.13%，小幅修改及接受稿件共 51.45%。而在 466 筆評語中，相較於評審報告份數百分比，大幅修改及小幅修改稿件的評語筆數的百分比均有增加，而在拒絕及接受稿件則為下降，接受稿件下降的幅度較大，可推論接受稿件的評語平均筆數最小。

表 4-1-4 評審報告份數及評語筆數分析－依評審者之出版建議

	評審者之出版建議									
	拒絕		大幅修改		小幅修改		接受		總數	
	數目	%	數目	%	數目	%	數目	%	數目	%
評審報告份數	20	19.42	30	29.13	35	33.98	18	17.47	103	100.00
評語筆數	89	19.10	154	33.05	162	34.76	61	13.09	466	100.00

另外就評審報告的字數來看，全樣本由 40 字至 4940 字不等，平均字數為 703 字（若刪除兩極端各 5%後為 629 字），中位數及標準差分別為 556 字及 653 字，呈右偏態分布，四分位全距為 576 字（見表 4-1-5 及附錄二、評審報告之總字數統計量分析）。

若區分量化研究與非量化研究，前者之評審報告字數由 66 字至 2928 字，平均數為 549 字（刪除兩極端各 5%後為 469 字），中位數為 392 字；標準差為 552 字，四分位全距為 498 字；非量化研究的評審報告字數則由 40 字至 4940 字，平均字數為 813 字（刪除兩極端各 5%後為 743 字），中位數為 663 字，標準差為 701 字，四分位全距為 645 字。綜合來看，量化研究的評審報告字數的分布較非量化研究為集中，前者主要在 200 至 400 字之間，而量化研究集中於 400 字至 700 字之間（見附錄二、評審報告之總字數統計量分析）。

若就稿件領域來看，社會科學研究之評審報告字數由 66 字至 2928 字，平均數為 634 字（刪除兩極端各 5%後為 562 字），中位數為 416 字；標準差為 615 字，四分位全距為 573 字；人文科學研究的評審報告字數則由 40 字至 4940 字，平均字數為 752 字（刪除兩極端各 5%後為 685 字），中位數為 619 字，標準差為 680

doi:10.6342/NTU201602255

字，四分位全距為 539 字。綜合來看，社會科學研究之評審報告字數分布較人文科學研究略為集中，前者主要在 100 至 400 字之間，而人文科學研究則集中於 300 字至 600 字之間（見附錄二、評審報告之總字數統計量分析）。

表 4-1-5 評審報告字數統計量比較—全樣本、稿件研究類型及稿件領域（*n*=103）

總字數		全樣本	稿件研究類型		稿件領域	
			量化研究	非量化研究	社會科學	人文科學
個數	有效的	103	43	60	43	60
	遺漏值	0	0	0	0	0
平均數		703.02	549.16	813.28	633.84	752.60
平均數的95%信賴區間	下限	575.36	379.19	632.30	444.56	576.94
	上限	830.68	719.16	994.27	823.11	928.26
刪除兩極端各5%觀察值之平均數		629.14	469.56	743.76	562.31	685.89
中位數		556.00	392.00	663.00	416.00	619.00
眾數		157[a]	201[a]	157[a]	201[a]	157[a]
標準差		653.210	552.368	700.601	615.020	680.004
範圍		4900	2862	4900	2862	4900
最小值		40	66	40	66	40
最大值		4940	2928	4940	2928	4940
總和		72411	23614	48797	633.84	752.60
百分位數	25	320.00	201.00	410.50	201.00	392.75
	50	556.00	392.00	663.00	416.00	619.00
	75	896.00	699.00	1055.25	774.00	931.75

a. 存在多個眾數，顯示的為最小值

第二節　評語筆數及字數分析

本研究由 103 份評審報告中共分析出 466 筆評語（可重複分類），就全樣本分析，每份評審報告之評語筆數由 1 至 9 筆不等，平均筆數為 4.52 筆，中位數為 4 筆，標準差為 1.754 筆，呈右偏態分布且接近常態形狀（見表 4-2-1 及附錄三、評

doi:10.6342/NTU201602255

審者之評語筆數統計量分析）。

若區分量化研究與非量化研究，前者的平均筆數為 4.88 筆（刪除兩極端各 5%觀察值之平均數各分別為 4.84），後者為 4.27 筆（刪除兩極端各 5%觀察值之平均數各分別為 4.22 筆），而量化研究的標準差（1.991 筆）及四分位全距（3 筆）均大於非量化研究（標準差為 1.528 筆；四分位全距為 2 筆），因此量化研究之評語筆數的分布較非量化研究為分散（見表 4-2-1 及附錄三、評審者之評語筆數統計量分析）。

就領域來看，社會科學研究的平均筆數為 5.14 筆（刪除兩極端各 5%觀察值之平均數各分別為 5.10），人文科學研究為 4.08 筆（刪除兩極端各 5%觀察值之平均數各分別為 4.04 筆），而社會科學研究的標準差（1.971 筆）及四分位全距（3 筆）均大於人文科學研究（標準差為 1.441 筆；四分位全距為 2 筆），因此社會科學研究之評語筆數的分布較人文科學研究為分散（見表 4-2-1 及附錄三、評審者之評語筆數統計量分析）。

doi:10.6342/NTU201602255

表 4-2-1 評語筆數統計量比較—依全樣本、稿件研究類型及稿件領域（*n*=466）

評語筆數		全樣本	稿件研究類型		稿件領域	
			量化研究	非量化研究	社會科學	人文科學
個數	有效的	103	43	60	43	60
	遺漏值	0	0	0	0	0
平均數		4.52	4.88	4.27	5.14	4.08
平均數的95%信賴區間	下限	4.18	4.27	3.87	4.53	3.71
	上限	4.87	5.50	4.66	5.75	4.46
刪除兩極端各5%觀察值之平均數		4.46	4.84	4.22	5.10	4.04
中位數		4.00	5.00	4.00	5.00	4.00
眾數		5	6	5	6	5
標準差		1.754	1.991	1.528	1.971	1.441
範圍		8	8	6	7	7
最小值		1	1	2	2	1
最大值		9	9	8	9	8
百分位數	25	3.00	3.00	3.00	3.00	3.00
	50	4.00	5.00	4.00	5.00	4.00
	75	6.00	6.00	5.00	6.00	5.00

進一步分析評語之平均筆數與出版建議的關係，就全數樣本來看，表 4-2-2 顯示大幅修改稿件的平均筆數最大有 5.13 筆，接受稿件評語最小為 3.39 筆。但是若區分量化研究與非量化研究，前者的拒絕稿件及大幅修改稿件之平均筆數較高，分別為 5.72 筆及 5.64 筆，接受稿件最低只有 3 筆；而非量化研究以小幅修改稿件（4.81 筆）及大幅修改稿件（4.69 筆）的平均筆數較高，拒絕稿件（3.93 筆）與接受稿件(3.5 筆)的平均筆數略低。另就稿件領域分析，社會科學研究以拒絕(6.20 筆）及大幅修改稿件（5.68 筆）的評語平均筆數較高，而接受稿件為最低，只有 2 筆；而人文科學研究的評語平均筆數較高者為小幅修改（4.67 筆）及大幅修改稿件（4.18 筆），而拒絕與接受稿件的差異有限，分別為 3.87 筆及 3.56 筆。

有關評語之平均字數與出版建議的關係，表 4-2-2 顯示在全數樣本中以大幅修

doi:10.6342/NTU201602255

改稿件的評語之平均字數最高為 208 字，而在拒絕稿件、小幅修改及接受稿件的差異不大。若就量化研究來看，接受稿件的評語之平均字數最高有 213 字，其次為拒絕稿件的 171 字；而非量化研究的評語之平均字數最高者為大幅修改稿件的 256 字，其次為拒絕稿件的 166 字，小幅修改稿件與接受稿件之字數相當。另就稿件領域分析，社會科學研究的評語平均字數以大幅修改稿件最高有 217 字，接受稿件最低有 145 字，而人文科學研究以大幅修改稿件之 188 字及拒絕稿件的 171 字較高，最低者為小幅修改稿件的 122 字。

表 4-2-2 評語之平均份數、筆數與字數分析（*n*=466）

	評審者之出版建議			
	拒絕	大幅修改	小幅修改	接受
樣本全數				
評審報告份數	20	30	35	18
評語平均筆數	4.45	5.13	4.63	3.39
評語平均字數	167.00	208.00	146.00	154.00
稿件研究類型				
量化研究				
評審報告份數	6	14	19	4
評語平均筆數	5.67	5.64	4.47	3.00
評語平均字數	171.00	162.00	151.00	213.00
非量化研究				
評審報告份數	14	16	16	14
評語平均筆數	3.93	4.69	4.81	3.50
評語平均字數	166.00	256.00	140.00	140.00
稿件領域				
社會科學	5	19	17	2
評審報告份數	5	19	17	2
評語平均筆數	6.20	5.68	4.59	2.00
評語平均字數	161	217	172	145
人文科學	15	11	18	16
評審報告份數	15	11	18	16
評語平均筆數	3.87	4.18	4.67	3.56
評語平均字數	171	188	122	155

doi:10.6342/NTU201602255

第三節　評語正負面向及評審標準分析

一、評語正負面向分析

許多研究指出評審者的意見多為正負夾雜，而且負面多於正面，尤其是拒絕稿件的負面評語較接受稿件為高（Gosden, 2003; Robson et al., 2015）。Bakanic 等人（1989）分析 *American Sociological Review* 的評審報告發現，無論是拒絕稿件或接受稿件都是負面評語多於正面評語，而且負面與正面評語比例在拒絕稿件為 5:1，接受稿件為 4:1。

本研究亦發現評審者的負面評語多於正面評語，在 466 筆評語中有 130 筆（28%）為正面評語，平均筆數為 1.26 筆；負面評語則有 336 筆（72%），平均筆數為 3.26 筆，正負評語之比例為 1:2.58。若進一步就出版建議來看，表 4-3-1 顯示在拒絕稿件、大幅修改，以及小幅修改稿件都是負面評語的比例較高，拒絕稿件的正負評語比例更是高達 1:13.83，不過接受稿件的正面評語卻微幅高於負面評語為 1:0.85。此一結果與 Robson 等人（2015）的研究較為接近，根據 *Journal of Advertising Research* 的評審報告發現，拒絕或大幅修改稿件主要是負面評語，接受稿件則多為正面，至於小幅修改稿件則是正負參半。

此外本研究亦發現量化研究與非量化研究在拒絕、大幅修改，以及小幅修改稿件都是負面評語較多，只有接受稿件是正面評語較多。在量化研究部份，拒絕稿件之正負評語比例高達 1:32.33、大幅修改稿件為 1:9，接受稿件的正面評語為負面評語的 2 倍。而非量化研究方面，拒絕稿件之正負評語比例為 1:10.11，接受稿件的正面評語僅略領先負面評語。若就稿件領域分析，社會科學研究的拒絕稿件之正負評語比例最高為 1:30.00，接受稿件的正面評語為負面評語的 3 倍；至於人文科學研究在不同出版建議之正負評語比例與非量化研究類似，但是在大幅修改的負面評語比例增加將近一倍（見 4-3-1）。

doi:10.6342/NTU201602255

表 4-3-1 評語平均筆數之正負面向及正負比分析（*n*=466）

	評審者之出版建議				
	拒絕	大幅修改	小幅修改	接受	總數
評語平均筆數	**4.45**	**5.13**	**4.63**	**3.39**	**4.52**
正面平均筆數	0.30	0.96	1.77	1.83	1.26
負面平均筆數	4.15	4.17	2.86	1.56	3.26
評語平均筆數正負比	**1:13.83**	**1:4.31**	**1:1.62**	**1:0.85**	**1: 2.58**
稿件研究類型					
量化研究	1:32.33	1:9.00	1:1.56	1:0.49	1:3.20
非量化研究	1:10.11	1:2.57	1:1.63	1:0.96	1:2.20
稿件領域					
社會科學	1:30.00	1:4.14	1:1.60	1:0.33	1:3.20
人文科學	1:10.60	1:4.75	1:1.63	1:0.90	1:2.27

若就評語之平均字數來看，Robson 等人（2015）的廣告學期刊研究指出，大幅修改稿件的平均字數最多，其次為拒絕及小幅修改，接受稿件的字數最少。本研究若就全數樣本分析，表 4-3-2 顯示平均字數最高者為大幅修改（208 字），但是其他三種稿件的字數差異不大；另拒絕稿件的負面評語平均字數約為正面稿件的 2 倍，大幅修改或小幅修改的正負比例相尚，而接受稿件的正面評語平均字數為負面的 1.5 倍強。

以量化研究來看，拒絕稿件的正面評語平均字數為負面評語之 35 倍強，而大幅修改及小幅修改之正負比相當，接受稿件的正面評語平均字數為負面評語的 1.6 倍；相較量化研究，非量化研究在各個出版建議之評語平均字數之正負比較小，惟拒絕大幅修改稿件是負面評語的平均字數較多，而在小幅修改及接受稿件則是正面評語的平均字數較高。另就稿件領域來看，社會科學研究之拒絕稿件的負面評語平均字數為正面評語的 33 倍強，但是在大幅修改稿件卻是正面評語平均字數較多，小幅修改的負面評語平均字數略高於正面，接受稿件之正面評語平均字數為負面評語約 4 倍；而在人文科學研究以拒絕稿件及大幅修改稿件都是負面評語平均字數均略高；至於小幅修改及接受稿件則是正面評語平均字數較高（見表

doi:10.6342/NTU201602255

4-3-2）。整體來看，相較於非量化研究及人文科學研究，量化研究與社會科學研究的評審者在拒絕稿件時，通常使用較多文字說明拒絕稿件的理由。

表 4-3-2 評語平均字數正負面向及正負比分析（*n*=466）

	評審者之出版建議				
	拒絕	大幅修改	小幅修改	接受	總數
評語之平均字數	**167**	**208**	**146**	**154**	**172**
正面平均字數	93	209	149	184	169
負面平均字數	173	208	144	119	173
評語平均字數正負比	**1:1.86**	**1:0.99**	**1:0.97**	**1:0.65**	**1:1.02**
稿件研究類型					
量化研究	1:35.12	1:1.36	1:1.06	1:0.61	1:1.05
非量化研究	1:1.55	1:1.09	1:0.86	1:0.70	1:1.02
稿件領域					
社會科學	1:33.23	1:0.97	1:1.06	1:0.24	1:1.05
人文科學	1:1.60	1:1.12	1:0.85	1:0.66	1:0.97

另 Bakanic 等人（1989）以社會學期刊進行的研究指出：「沒有一篇稿件全部都是正面評語，只是有些稿件的負面評語較多（p. 639）。」表 4-3-3 顯示本研究的 103 份評審報告中有超過 1/3（39 份，38%）的評語為全正或全負，其中有 98%（37 份評審報告）為全負評語，全正評語的評審報告只有 2 份。若就稿件出版建議分析，全正評語者都是接受稿件，進一步以稿件研究類型來看，量化研究與非量化研究各一篇；而若以稿件領域來看，社會科學與人文稿學亦各有一篇。

至於全負評語之評審報告，大部份（64.86%）屬於拒絕稿件及大幅修改稿件。但是有超過 1/3（35.14%）為接受及小幅修改稿件，出現評審者評語與出版建議不一致的情況，亦即評審者在建議小幅修改或接受稿件時，都是使用負面評語。若就稿件研究類型分析，量化研究全負評語之比例略大於非量化研究，分別為 37.5% 及 33.33%，而社會科學研究的全負評語為 38.46%，亦高於人文科學研究的 33.33%（見表 4-3-3）。

doi:10.6342/NTU201602255

表 4-3-3 評審報告之評語為全正及全負分析（*n*=39）

評審報告	評審者之出版建議（份數與%）				
	拒絕	大幅修改	小幅修改	接受	總數
全部正面評語	0/0.00	0/0.00	0/0.00	2/100.00	2/5.13
稿件研究類型					
量化研究	0/0.00	0/0.00	0/0.00	1/100.00	1/50.00
非量化研究	0/0.00	0/0.00	0/0.00	1/100.00	1/50.00
稿件領域					
社會研究	0/0.00	0/0.00	0/0.00	1/100.00	1/50.00
人文研究	0/0.00	0/0.00	0/0.00	1/100.00	1/50.00
全部負面評語	15/40.54	9/24.32	11/29.73	2/5.41	37/94.87
稿件研究類型					
量化研究	4/25.00	6/37.5	6/37.5	0/0.00	16/43.24
非量化研究	11/52.38	3/14.29	5/23.81	2/9.52	21/56.76
稿件領域					
社會科學	4/30.77	4/30.77	5/38.46	0/0.00	13/35.14
人文科學	11/45.83	5/20.83	6/25.00	2/8.33	24/64.86

二、評審標準分析

本研究的稿件科學品質分析是採用 Bornmann、Nast 等人（2008）提出之期刊同儕審查評審標準分類架構，該架構共有 9 大類，103 份評審報告在編碼完成後只出現 6 大類，包括：貢獻度相關、寫作與呈現、設計與概念、方法與統計、結果討論，以及文獻分析，至於理論、作者聲望或所屬機構，以及道德 3 大類則未有相關評語；亦即我國社會暨人文科學領域研究中有關理論的論述較少，且未出現有違研究道德的情況，此外評審者並沒有將作者聲望或所屬機構列為稿件的審查標準。

許多研究指出寫作與呈現是評審者經常使用的標準，而且有一半以上為負面評語（Bornmann, Nast, et al., 2008），本研究亦發現寫作與呈現無論在全樣本、或區分稿件研究類型（量化研究或非量化研究），或區分稿件領域（社會科學或人文科學），都是經常使用的評語，排名為第一或第二；而且此類評語的負面比例甚高，

doi:10.6342/NTU201602255

約在 80%上下。Weller（2002）的回顧專書指出，大部份寫作與呈現的批評都是可以修正的意見，本研究亦呼應此一說法。

除了寫作與呈現的評語外，就全數樣本來看，表 4-3-4 顯示最常使用的標準為結果討論及貢獻度相關，而在量化研究中，貢獻度相關及方法與統計兩項標準排名居前，其次為結果討論；至於非量化研究中最常使用者為結果討論，其次為貢獻度相關及設計與概念，方法與統計的評語則闕如。另就稿件領域分析，表 4-3-4 顯示社會科學研究最常使用的評審標準為貢獻度相關及結果討論，與非量化研究相比，方法與統計的排名與百分比均略為下降，主要因為社會科學研究中有將近 1/3 的非量化研究，未有此類評語；而人文科學研究的前三名為結果討論、貢獻度相關，以及設計與概念。

表 4-3-4 評審標準使用情形與排名，並與國外文獻進行比較—全樣本與稿件研究類型（*n*=466）

	本研究之評審標準（使用情形%與排名）			Bornmann 與 Nast 等人（2008）文獻排序[a]	
評審標準	全樣本	量化研究	非量化研究	頻率排名	重要性排名
貢獻度相關	19.96/(3)	8.37/(2)	11.59/(3)	1	5
寫作與呈現	24.46/(1)	10.52/(1)	13.95/(2)	2	6
設計與概念	14.81/(4)	5.15/(5)	9.66/(4)	3	3
方法與統計	8.37/(6)	8.37/(2)	0.00	4	4
結果討論	22.32/(2)	7.30/(4)	15.02/(1)	5	1
文獻分析	10.09/(5)	4.94/(6)	5.15/(5)	6	7
理論	0.00	0.00	0.00	7	2
作者聲望或所屬機構	0.00	0.00	0.00	8	9
道德	0.00	0.00	0.00	9	8
總數	100.00	44.64	55.36	-	-

a. 資料來源：“Do editors and referees look for signs of scientific misconduct when reviewing manuscripts?” by L. Bornmann, I. Nast & H.-D. Daniel, 2008, *Scientometrics*, *77*(3), pp. 415-432.

doi:10.6342/NTU201602255

另表 4-3-4 亦顯示量化研究的評審標準排序與 Bornmann 與 Nast 等人（2008）的頻率排名接近，除了寫作與呈現之外，兩者都同樣重視貢獻度相關及方法與統計，比較大的差異在設計與概念。在 Bornmann 與 Nast 等人的後設分析文獻中，無論就頻率或重要性來看，設計與概念都名列第 3，但是本研究卻僅排名第 5。研究者推論可能因為目前期刊同儕審查的研究文獻，以醫學領域開始較早，文獻亦眾（Rennie, 2003; Weller, 2002），此一領域之實驗性研究較多，對於設計與概念的重視反應在 Bornmann 與 Nast 等人的研究結果。另就社會科學與人文科學的評審標準排名與 Bornmann 與 Nast 等人（2008）的排名相比，表 4-3-5 顯示兩者之間的差異甚大。

表 4-3-5 評審標準使用情形與排名，並與國外文獻進行比較－全樣本與稿件領域（*n*=466）

本研究之評審標準（使用情形%與排名）				Bornmann 與 Nast 等人（2008）文獻排序 [a]	
評審標準	全樣本	社會科學	人文科學	頻率排名	重要性排名
貢獻度相關	19.96/(3)	9.66/(2)	10.30/(3)	1	5
寫作與呈現	24.46/(1)	12.45/(1)	12.02/(2)	2	6
設計與概念	14.81/(4)	5.58/(6)	9.23/(4)	3	3
方法與統計	8.37/(6)	6.01/(4)	2.36/(6)	4	4
結果討論	22.32/(2)	7.94/(3)	14.38/(1)	5	1
文獻分析	10.09/(5)	5.79/(5)	4.29/(5)	6	7
理論	0.00	0.00	0.00	7	2
作者聲望或所屬機構	0.00	0.00	0.00	8	9
道德	0.00	0.00	0.00	9	8
總數	100.00	47.42	52.58	-	-

a.資料來源：同表 4-3-4

若進一步分析評審標準與出版建議的關係，表 4-3-6 顯示除了寫作與呈現外，在量化研究中拒絕稿件最重要的負面評語為方法與統計及貢獻度相關，在接受稿

doi:10.6342/NTU201602255

件中最重要之正面評語為貢獻度相關；至於非量化研究之拒絕稿件的主要負面評語為結果討論及設計與概念,接受稿件之主要正面評語為貢獻度相關及結果討論，不過在主要負面評語中亦包括結果討論在內，排名僅次於寫作與呈現；而在小幅修改稿件及大幅修改稿件中，結果討論的批評均超過寫作與呈現而是排名第一的負面意見。

表 4-3-6 評審者使用之評語與評審者之出版建議分析—依稿件研究類型（*n*=466）

量化研究								
	評審者之出版建議							
評審標準	拒絕		大幅修改		小幅修改		接受	
(%)	正面	負面	正面	負面	正面	負面	正面	負面
貢獻度相關	3	18	5	11	13	6	33	0
寫作與呈現	0	12	5	23	8	18	0	8
設計與概念	0	15	0	11	1	11	0	8
方法與統計	0	24	0	16	8	9	17	8
結果討論	0	15	0	15	5	12	17	8
文獻分析	0	15	0	13	4	6	0	0
總計	3	97	10	90	39	61	67	33
非量化研究								
	評審者之出版建議							
評審標準	拒絕		大幅修改		小幅修改		接受	
(%)	正面	負面	正面	負面	正面	負面	正面	負面
貢獻度相關	4	15	12	12	12	9	16	2
寫作與呈現	4	20	5	17	8	18	6	24
設計與概念	2	20	4	12	9	6	12	4
方法與統計	0	0	0	0	0	1	0	0
結果討論	0	25	3	20	9	19	16	18
文獻分析	0	11	4	11	0	9	0	0
總計	9	91	28	72	38	62	51	49

doi:10.6342/NTU201602255

表若 4-3-7 以稿件領域進行分析，社會科學研究中拒絕稿件最重要的負面評語為貢獻度相關及方法與統計，在接受稿件中最重要之正面評語為貢獻度相關及設計與概念，而負面評語中以結果討論最高；至於人文科學研究之拒絕稿件的主要負面評語為結果討論及設計與概念，接受稿件之主要正面評語為貢獻度相關及結果討論，不過在主要負面評語中亦包括結果討論在內（占 16%）。

表 4-3-7 評審者使用之評語與評審者之出版建議分析—稿件領域（*n*=466）

社會科學								
	評審者之出版建議							
評審標準	拒絕		大幅修改		小幅修改		接受	
(%)	正面	負面	正面	負面	正面	負面	正面	負面
貢獻度相關	3	19	9	12	13	5	25	0
寫作與呈現	0	13	5	23	8	22	25	0
設計與概念	0	13	2	10	1	9	25	0
方法與統計	0	19	0	10	8	6	0	0
結果討論	0	16	1	15	5	13	0	25
文獻分析	0	16	3	10	4	6	0	0
總計	3	97	19	81	38	62	75	25
人文科學								
	評審者之出版建議							
評審標準	拒絕		大幅修改		小幅修改		接受	
(%)	正面	負面	正面	負面	正面	負面	正面	負面
貢獻度相關	3	14	7	11	12	10	19	2
寫作與呈現	3	19	7	13	8	14	4	23
設計與概念	2	21	2	15	8	8	9	5
方法與統計	0	3	0	4	1	4	4	2
結果討論	0	24	2	24	8	18	18	16
文獻分析	0	10	0	15	0	8	0	0
總計	9	91	17	83	38	62	53	47

doi:10.6342/NTU201602255

第四節　評審信度分析

同儕審查並非一種隨機過程，理論上評審者之間的判斷應該會有某種程度的一致性，亦即存在合理的評審信度，但是根據目前的研究結果發現評審者之間的信度不高（Bornmann, Mutz, et al., 2010），許多學者因而質疑同儕審查的公平性（Cicchetti, 1991; Bornmann, 2011a; Marsh et al., 2007），但是有學者認為低信度表示評審者可以誠實地表達不同的意見，也回應學術界對於研究前沿的不同看法（Cole, J. R., 2000; Cole, S., 1992）。

本研究的 48 篇稿件都至少有 2 位評審者，表 4-4-1 依照評審者擲還稿件日期的先後區分為第一位評審者及第二位評審者，前者建議接受稿件者占 8.86%、後者占 5.36%；而拒絕稿件在第一位評審者有 7.69%、第二位評審者有 12.35%，整體來看，似乎較早擲還的評審者建議接受稿件者相對較多，而拒絕稿件相對較少，另本研究樣本中有第三位評審者不多，故不予計入。

表 4-4-1 第一位評審者及第二位評審者之出版建議分析（*n*=103）

	評審者之出版建議（%）			
	拒絕	大幅修改	小幅修改	接受
第一位評審者	7.69	17.48	16.08	8.86
第二位評審者	12.35	14.69	17.48	5.36
總計	20.05	32.17	33.57	14.22

（一）出版建議的評審信度

目前的評審信度研究大多針對評審者的出版建議進行分析（Weller, 2002），Bornmann、Mutz等人（2010）的後設分析文獻指出，期刊評審信度的Cohen's Kappa平均值為0.17，而且若是在文獻中未清楚說明評分系統（rating system）者之評審信度較高。Cicchetti（1991）認為同儕審查的評分系統乃評審信度研究的必備資訊，未提供此類訊息之文獻的研究品質堪慮。

doi:10.6342/NTU201602255

本研究所蒐集之3種期刊的出版建議系統略有差異，經綜整為4類為：拒絕稿件、大幅修改（再審或接受）稿件、小幅修改（接受）稿件，以及接受稿件。為利與其他文獻進行比較，本研究將出版建議之評審信度分析分為3種模式，其一為模型A：出版建議分為4類，包括拒絕、大幅修改、小幅修改，以及接受稿件；其二為模型B：出版建議分為3類，包括拒絕、大幅修改（再審或接受），以及接受（含小幅修改）稿件；其三為模型C：出版建議分為2類，包括拒絕（含大幅修改）與接受（含小幅修改）稿件。

表4-4-2顯示模型A的Cohen's Kappa值僅0.10；模型B及模型C的Cohen's Kappa值都為0.23，但是模型C在Actual agreement上高於模型B。Daniel（1993, p. 23）說明Kappa係數的意義指出：「Kappa係數0.23表示評審之間對於23%的文稿內容之評價趨於一致，而不只是隨機的結果。」Fleiss（1981）認為Kappa值在0.4以下為信度低，0.4至0.6之間為一般。

表 4-4-2 出版建議之評審信度分析（*n*=103）

出版建議	二位評審者同時回應之數	Actual Agreement	Chance Agreement	Cohen's Kappa Coefficient
模型 A	104	0.34	0.27	0.10
拒絕	6	0.83	0.83	0.00
大幅修改	19	0.79	0.66	0.38
小幅修改	6	0.67	0.44	0.40
接受	4	1.00	1..00	1.00
模型 B	104	0.51	0.36	0.23
拒絕	6	0.83	0.83	0.00
大幅修改	19	0.79	0.66	0.38
接受(含小幅修改)	28	0.71	0.51	0.42
模型 C	104	0.62	0.50	0.23
拒絕(含大幅修改)	36	0.83	0.75	0.32
接受(含小幅修改)	28	0.71	0.51	0.42

有些研究指出期刊評審者對於拒絕稿件的一致性較高（Cicchetti, 1991, 1997; Daniel, 1993; Ingelfinger, 1974），而且是接受稿件的2倍（Weller, 2002）。本研究並

doi:10.6342/NTU201602255

未支持此一看法，無論是模型A、模型B或模型C，拒絕稿件的Cohen's Kappa值均不高，反而是接受稿件的Kappa值較高，三種模式分別為0.40、0.42，以及0.42（見表4-4-2）。若進一步區分稿件研究類型之出版建議評審信度，表4-4-3顯示量化研究之3種模式的Cohen's Kappa值都小於0；非量化研究之3種模式的Cohen's Kappa值分別為0.23、0.46、0.44。而若以稿件領域分析，社會科學研究之3種模式的Cohen's Kappa值都小於0；人文科學研究之3種模式的Cohen's Kappa值分別為0.11、0.34、0.32。整體來看非量化研究與人文科學研究的出版建議之評審信度均較量化研究與社會科學研究為高。

表 4-4-3 稿件研究類型及稿件領域之出版建議評審信度分析（*n*=103）

出版建議	二位評審者同時回應之數	Actual Agreement	Chance Agreement	Cohen's kappa Coefficient
稿件研究類型				
量化研究				
模式 A[1]	39	0.21	0.32	<0
模式 B[2]	39	0.25	0.36	<0
模式 C[3]	39	0.45	0.52	<0
非量化研究				
模式 A[1]	65	0.42	0.25	0.23
模式 B[2]	65	0.67	0.37	0.46
模式 C[3]	65	0.72	0.50	0.44
稿件領域				
社會科學				
模式 A[1]	44	0.34	0.37	<0
模式 B[2]	44	0.36	0.40	<0
模式 C[3]	44	0.55	0.58	<0
人文科學				
模式 A[1]	60	0.33	0.26	0.11
模式 B[2]	60	0.62	0.42	0.34
模式 C[3]	60	0.67	0.51	0.32

備註：1. 模式 A（4 類）：拒絕、大幅修改、小幅修改，以及拒絕；2. 模式 B（3 類）：拒絕、大幅修改，以及接受（含小幅修改）；3. 模式 C（2 類）：拒絕（含大幅修改）及接受（含小幅修改）。

doi:10.6342/NTU201602255

（二）評審標準的評審信度

表 4-4-4 針對評審者使用的評審標準進行信度分析，發現量化研究評審標準的 Cohen's Kappa 為 0.22，非量化研究則略高為 0.33。若就個別評審標準來看，量化研究在貢獻度相關、結果討論，以及文獻分析 3 項標準之 Cohen's Kappa 為 1；而在寫作與呈現、設計與概念，以及方法與統計的 Cohen's Kappa 為 0 或小於 0。而在非量化研究中，文獻分析的評審信度最高為 0.5，其次為結果討論（0.39）、貢獻度相關（0.31），評審者對於寫作與呈現（0.24）及設計與概念（0.14）的看法較不一致。一個有趣的現象是儘管寫作與呈現是經常使用的標準，但是評審信度並不高，在量化研究為 0；非量化研究也僅 0.24。

表 4-4-4 評審標準信度分析表—依稿件研究類型（*n*=103）

評審標準	二位評審者同時回應之數	Actual Agreement	Chance Agreement	Cohen's Kappa Coefficient
量化研究				
貢獻度相關	5	1.00	0.68	1.00
寫作與呈現	6	0.83	0.83	0.00
設計與概念	6	0.83	0.83	0.00
方法與統計	10	0.70	0.74	<0.00
結果討論	8	1.00	1.00	1.00
文獻分析	5	1.00	1.00	1.00
總數	40	0.87	0.84	0.22
非量化研究				
貢獻度相關	9	0.67	0.52	0.31
寫作與呈現	13	0.69	0.60	0.24
設計與概念	10	0.50	0.42	0.14
結果討論	28	0.75	0.59	0.39
文獻分析	4	0.75	0.50	0.50
總數	64	0.69	0.54	0.33

表 4-4-5 顯示社會科學研究評審標準的 Cohen's Kappa 為 0.32，人文科學研究則略高為 0.35。若就個別評審標準來看，社會科學研究在貢獻度相關及結果討論之 Cohen's Kappa 為 1，文獻分析為 0.59；而在寫作與呈現、設計與概念，以及方

doi:10.6342/NTU201602255

法與統計的 Cohen's Kappa 為 0 或小於 0。另在人文科學研究，文獻分析的評審信度最高為 1，其次為結果討論（0.39）、寫作與呈現（0.30），貢獻度相關及設計與概念都是 0.25。寫作與呈現的評審信度在社會科學小於 0，人文科學研究較高為 0.30。

表 4-4-5 評審標準信度分析表—依稿件領域（*n*=103）

評審標準	二位評審者同時回應之數	Actual Agreement	Chance Agreement	Cohen's Kappa Coefficient
社會科學研究				
貢獻度相	6	1.00	0.56	1.00
寫作與呈	8	0.00	0.50	<0.00
設計與概念	8	0.63	0.63	0.00
方法與統計	7	0.71	0.76	<0.00
結果討論	8	1.00	1.00	1.00
文獻分析	7	0.86	0.65	0.59
總數	44	0.82	0.73	0.32
人文科學研究				
貢獻度相	8	0.63	0.50	0.25
寫作與呈	11	0.73	0.61	0.30
設計與概念	8	0.63	0.50	0.25
方法與統計	3	0.67	0.67	0.00
結果討論	28	0.75	0.59	0.39
文獻分析	2	1.00	1.00	1.00
總數	60	0.72	0.56	0.35

doi:10.6342/NTU201602255

第五節　評審的公平性及課責性

一、評審的公平性

同儕審查是學術社群的自我規範機制，公平性則是同儕審查機制合理性的基礎。Bornmann（2011b）認為同儕審查的公平性研究是為了提升審查作業的公平性並減少評審偏見，也就是所謂的興利除弊。有學者認為公平性是基於科學規範的普遍主義，科學的評價必須基於科學標準，而不應受到非科學品質因素的影響，例如個人的性別、服務機構或社會地位等（Merton, 1942; Ziman, 2000）。

目前少數以稿件評審報告進行的研究中，已有學者證實某些管理學及物理學期刊，作者的知名度及所屬機構影響評審結果（Beyer et al., 1995; Zuckerman & Merton, 1971）；不過 Gilliland 與 Cortina（1997）分析 *Journal of Applied Psychology* 的評審報告發現，評審者所使用的評語均與科學品質相關。本研究為探討評審的公平性亦針對非科學品質評語進行編碼，發現評審者只針對稿件的科學品質進行審查，並未出現有關作者個人特質或所屬機構等非科學品質之評語。

有學者認為評審者的審查習性與風格可能影響評審的公平性，Spencer 等人（1986）以心理學期刊的評審報告進行研究，發現使用情緒性或不確定評論超過 25%，推論可能受到評審者個人偏見的影響。表 4-5-1 顯示有少數評審者使用情緒性評語（5.83%），大多屬於非量化研究或人文科學，而且都是拒絕稿件或大幅修改稿件，接受稿件及小幅修改稿件並未有情緒性評語。本研究之情緒性用語是否為為評審者的個人偏見尚難論斷，不過可能與評審者的寬嚴態度或個人風格相關，需要進一步研究。

doi:10.6342/NTU201602255

表 4-5-1 評審者情緒性評語分析（*n*=6）

	評審者之出版建議			
	拒絕	大幅修改	小幅修改	接受
稿件研究類型				
量化研究	0	1	0	0
非量化研究	4	1	0	0
稿件領域				
社會科學	0	1	0	0
人文科學	4	1	0	0

評審者向作者透露出版建議也是期刊編輯關心的議題，因為會造成審查作業的困擾（Lovejoy et al., 2011），有學者分析 *Journal of International Business Studies* 於 2011 至 2012 年的最佳評審獲獎者的審查報告，整理出 6 項共通點，其中之一為在致作者的評語中從未出現關於出版建議的明示或暗示（Caligiuri & Thomas, 2013）。

此外 Smith (2006)認為有 1/4 至 1/3 的稿件，評審者可以確認作者身分，例如自我引用可能透露作者身份，而特定領域的長期研究者也極易辨識。研究者認為我國為小型科研國家，各學科領域的研究者不多，評審者可輕易推斷作者身份，透露出版建議之舉有可能存有潛在公平性問題。表 4-5-2 顯示在 103 份評審報告中有 11 份（10.67%）向作者透露出版建議，而且拒絕稿件與接受稿件都有，不過大部分屬於非量化研究或人文科學稿件。

表 4-5-2 評審者的評語透露出版建議（*n*=11）

	評審者之出版建議				
	拒絕	大幅修改	小幅修改	接受	總數
稿件研究類型					
量化研究	0	1	0	0	1
非量化研究	4	0	3	3	10
稿件領域					
社會科學	0	1	0	0	1
人文科學	4	0	3	3	10

doi:10.6342/NTU201602255

二、評審的課責性

在期刊同儕審查的作業中，評審者的主要功能有二，其一為協助期刊編輯決定接受或拒絕稿件，其二為協助作者改進稿件品質（van Rooyen et al., 1999）。因此評審者的評語應清楚表達看法，而且需與出版建議一致，以提供期刊編輯有價值的參考資訊，也避免讓稿件作者產生錯誤的期待。近年來為了提升期刊同儕審查的作業品質，個別期刊或相關學者提出各種看法，有建議設計評審者評鑑系統，由期刊編輯為評審者的表現打分數；有主張編訂評審指南，說明評審者的職能與責任；有認為舉辦同儕審查之評審訓練課程或研習會（Callhan & Schriger, 2002; De Vries et al., 2009; Schroter et al., 2004）。

本研究認為就同儕審查的作業管理層面來看，目前的評審公平性研究大多有其因果推論之侷限性，因此強化評審的課責性有其實質意義，並可具體提升同儕審查之效能。本研究之評審的課責性，強調評審者的評語與出版建議的一致性，也就是說評審者在建議接受稿件時，應具體指出稿件的優點，而在拒絕稿件時亦應提出難以改正之批評；因此將評審者的正面評語以「具體性」、負面評語以「可修改性」各分為 3 個等級進行評分。此一作法之目的不在為評審者打分數，而是讓評審者在評分時更加重視其評語與出版建議的合理性，也讓期刊編輯更加瞭解評審者的評語與出版建議之間應具有之關聯性。

（一）評語分數的啓發

本研究認為評審者在建議接受稿件時，應具體指出稿件科學研究價值，因此將評審者的正面評語以「具體性」（評語內容明確並專指）分為 3 級：不具體（1）、具體（2），以及很具體（3）；而評審者在建議拒絕稿件時亦應指出與稿件學品質相關且難以改正之批評，而將負面評語以「可修改性」（評語內容之修改難易程度）分為 3 級：困難修改（-3）、尚可（-2），以及容易修改（-1），以下就正負面向評語的平均分數進行分析，並討論不同等級評分與出版建議的關係。

doi:10.6342/NTU201602255

表 4-5-3 顯示全數樣本的正面評語平均分數為 1.82 分，其中接受稿件的分數最高（2.24 分），其次為拒絕稿件 1.83 分，小幅修改及大幅修改則分別是 1.79 分及 1.38 分。若就量化研究來看，其正面評語平均分數由接受稿件（2.75 分）到拒絕稿件(1 分)呈現下降的趨勢;但是非量化研究的正面評語平均分數則是接受稿件、拒絕稿件，以及小幅修改稿件的平均分數相當，在 2.00 上下；大幅修正稿件的 1.38 分為最低。就稿件領域分析，社會科學研究的拒絕稿件與接受稿件較低，分別為 1.00 及 1.33 分，小幅修改最高為 1.63 分；人文科學研究則是接受稿件、拒絕稿件，以及小幅修改稿件均高，分別為 2.33 分、2.00 分，以及 1.94 分，大幅修改稿件最低為 1.13 分

另外負面評語的全樣本平均分數為-2.24 分，由最低的拒絕稿件（-2.55 分）逐漸上升至最高的接受稿件（-1.75 分）。非量化研究的平均分數也呈現相同趨勢，但是拒絕稿件（-2.7 分）與接受稿件（-1.71 分）之間的差距超過 1 分。但是量化研究的負面平均分數則呈現不規則狀況，各類出版建議的平均分數均頗為接近在-2.3 分至-1.98 分之間，以拒絕稿件與大幅修改稿件的平均分數較低。若就稿件領域分析，社會科學研究之接受稿件的負面評語較低（-3.00 分），拒絕稿件居次（-2.30 分），小幅修改稿件最高（1.92 分），人文稿學研究之拒絕稿件分數最低（2.70 分），逐步上升至接受稿件之負面評語平均分數最高（1.70 分）。

doi:10.6342/NTU201602255

表 4-5-3 正面及負面評語平均分數與出版建議分析（*n*=466）

	評審者之出版建議				
	拒絕	大幅修改	小幅修改	接受	平均總計
正面評語平均分數	**1.83**	**1.38**	**1.79**	**2.24**	**1.82**
稿件研究類型					
量化研究	1.00	1.38	1.58	2.75	1.73
非量化研究	2.00	1.38	2.00	2.08	1.87
稿件研究領域					
社會科學	1.00	1.48	1.63	1.33	1.55
人文科學	2.00	1.13	1.94	2.33	2.01
負面評語平均分數	**-2.55**	**-2.30**	**-2.04**	**-1.75**	**-2.24**
稿件研究類型					
量化研究	-2.30	-2.21	-1.98	-2.00	-2.15
非量化研究	-2.70	-2.41	-2.10	-1.71	-2.32
稿件領域					
社會科學	-2.30	-2.18	-1.92	-3.00	-2.13
人文科學	-2.70	-2.55	-2.15	-1.70	-2.34

對於不同等級評分與出版建議之關係，就全數樣本來看，表 4-5-4 顯示正面評語中「很具體」(+3)評語多出現在接受稿件(63.64%)及小幅修正稿件(31.82%)，兩者合計達 95.46%，評語的具體性與出版建議的一致性甚高。研究者亦發現在拒絕稿件中有一筆「很具體」評語，經查係因所採分類架構之關係，評審者雖然肯定研究主題之創見，但是對於論述內容卻持否定的看法；至於「具體」(+2) 評語則以小幅修改稿件為最高（56.45%)，但是在「不具體」(+1) 評語中有超過一半（56.52%）屬於接受及小幅修正稿件，推估某些評審報告的正面評語可能不夠具體支持其出版建議。而在負面評語中「容易修改」(-1) 及「尚可修改」(-2) 的評語主要集中於小幅修改及大幅修改稿件，分別合計為 67.17%及 77.05%；「困難修改」(-3) 則大多屬於拒絕與大幅修改稿件（75.52%)，但是在小幅修改及接受稿件的比例亦有將近 1/4（24.49%)，推論某些評審報告的負面評語與拒絕稿件之出版建議可能存在不一致的情況。

若分析稿件研究類型，量化研究與非量化研究的整體的差異不大，但是仍有

doi:10.6342/NTU201602255

若干顯著不同。表 4-5-4 顯示就正面「很具體」(+3) 評語分析，量化研究在接受稿件有 75%；非量化研究僅 57.14%，也就是在非量化研究中比較可能出現正面評語不夠具體無法支持其接受稿件之出版建議。而在「困難修改」(-3) 評語的分布情況，量化研究與非量化研究則有同有異，相同之處在於兩者在接受稿件及小幅修改稿件都有 1/4 左右的「困難修改」(-3) 評語，量化研究為 22.8%，非量化研究為 25.56%，推估可能有評語與出版建議不一致的情況。至於相異之處為量化研究中的「困難修改」評語多出現在大幅修改稿件（49.12%），而非量化研究則主要在拒絕稿件（43.33%）(見表 4-5-4)，顯示兩種研究類型的評審者對於拒絕稿件的寬嚴程度或許有別，也可能受到期刊編輯政策或拒稿率的影響。

doi:10.6342/NTU201602255

表 4-5-4 各級評語評分與出版建議分析—全樣本及稿件研究類型（*n*=466）

	評審者之出版建議（%）			
	拒絕	大幅修改	小幅修改	接受
全數樣本				
正面評語				
不具體	4.35	39.13	43.48	13.04
具體	4.84	17.74	56.45	20.97
很具體	4.55	0.00	31.82	63.64
負面評語				
困難修改	37.42	38.10	20.41	4.08
尚可修改	15.57	40.98	36.07	7.38
容易修改	13.43	28.36	38.81	19.40
稿件研究類型				
量化研究				
正面評語				
不具體	4.55	22.73	72.73	0.00
具體	0.00	15.00	75.00	10.00
很具體	0.00	0.00	25.00	75.00
負面評語				
困難修改	28.07	49.12	21.05	1.75
尚可修改	16.67	41.67	38.89	2.78
容易修改	16.13	41.94	38.71	3.23
非量化研究				
正面評語				
不具體	4.26	54.17	16.67	25.00
具體	7.14	19.05	47.62	26.19
很具體	7.14	0.00	35.71	57.14
負面評語				
困難修改	43.33	31.11	20.00	5.56
尚可修改	14.00	40.00	32.00	14.00
容易修改	11.11	16.67	38.89	33.33

doi:10.6342/NTU201602255

就就稿件領域來看，在正面「很具體」(+3）評語分析，社會科學全部為小幅修改稿件；人文科學則有 70%為接受稿件；25%為小幅修改稿件。而在「困難修改」（-3）評語的分布情況，社會科學有 15.37%屬於小幅修改或接受稿件，人文科學則有 29.34%，推估可能有評語與出版建議不一致的情況。另外在社會科學的「困難修改」評語多出現在大幅修改稿件（36.84%），而人文科學研究則主要在拒絕稿件（44.57%）(見表 4-5-5)。顯示兩種研究類型的評審者對於拒絕稿件的寬嚴程度或許有別，也可能受到期刊編輯政策或拒稿率的影響。

表 4-5-5 各級評語評分與出版建議分析—稿件領域（*n*=466）

	評審者之出版建議（%）			
	拒絕	大幅修改	小幅修改	接受
社會科學研究				
正面評語				
不具體	3.70	40.74	48.15	7.41
具體	0.00	38.46	57.69	3.85
很具體	0.00	0.00	100.00	0.00
負面評語				
困難修改	25.45	58.18	14.55	1.82
尚可修改	14.10	50.00	35.90	0.00
容易修改	15.15	48.48	36.36	0.00
人文科學研究				
正面評語				
不具體	5.26	36.84	36.84	21.05
具體	8.33	2.78	55.56	33.33
很具體	5.00	0.00	25.00	70.00
負面評語				
困難修改	44.57	26.09	23.91	5.43
尚可修改	18.18	25.00	36.36	20.45
容易修改	11.76	8.82	41.18	38.24

doi:10.6342/NTU201602255

（二）評審的課責性－檢驗「困難修改」(-3）評語

拒絕稿件的理由一直是學者最為關切議題，Weller（2002）的回顧專書指出有些拒絕稿件的理由是可修改的，例如寫作與呈現方式、結論的詮釋、資料分析，以及文獻分析等。此外也有研究證實被某期刊拒絕之稿件，轉投其他期刊而獲刊的比例甚高（Alberts et al., 2008），而且在高影響力期刊所出版的文獻亦有許多是轉投稿件（Calcagno et al., 2012）。

綜合言之，表 4-5-6 顯示「困難修改」(-3）評語約有 1/4 出現在接受稿件或小幅修改稿件，研究者推論可能存在評語與出版建議不一致的情況；若就稿件研究類型分析，量化研究為 22.8%，非量化研究為 25.56%；若就稿件領域來看，社會科學研究為 16.37，人文科學研究為 29.34%。

表 4-5-6 困難修改(-3)評語與出版建議分析－依稿件研究類型與稿件領域(n=147)

	評審者之出版建議（評語個數與%）				
	拒絕	大幅修改	小幅修改	接受	總計
稿件研究類型					
量化研究	16(28.07%)	28(49.12%)	12(21.05%)	1(1.75%)	57(100.00%)
非量化研究	39(43.33%)	28(31.11%)	18(20.00%)	5(5.56%)	90(100.00%)
總計	55(37.41%)	56(38.10%)	30(20.41%)	6(4.08%)	147(100.00%)
稿件領域					
社會研究	14(25.45%)	32(58.18%)	8(14.55%)	1(1.82%)	55(100.00%)
人文研究	41(44.57%)	24(26.09%)	22(23.91%)	5(5.43%)	92(100.00%)
總計	55(37.41%)	56(38.10%)	30(20.41%)	6(4.08%)	147(100.00%)

此外，若換個角度思考，評審報告中沒有出現「困難修改」(-3）評語，評審者是否亦不宜給予拒絕稿件的建議。表 4-5-7 顯示評審報告中未出現「困難修改」(-3）評語、卻屬於拒絕稿件亦有 10%左右，無論是以稿件研究類型或稿件領域進行分析，比例差異不大。

doi:10.6342/NTU201602255

表 4-5-7 未出現困難修改（-3）評語與出版建議分析－依稿件研究類型與稿件領域（n=319）

	評審者之出版建議（評語個數與%）				
	拒絕	大幅修改	小幅修改	接受	總計
稿件研究類型					
量化研究	18(11.77%)	51(33.33%)	73(47.71%)	11(7.19%)	153(100.00%)
非量化研究	16(9.64%)	47(28.31%)	59(35.54%)	44(26.51%)	166(100.00%)
總計	34(10.66%)	98(30.72%)	132(41.38%)	55(17.24%)	319(100.00%)
稿件領域					
社會研究	17(10.24%)	76(45.78%)	70(42.17%)	3(1.81%)	166(100.00%)
人文研究	17(11.11%)	22(14.38%)	62(40.52%)	52(33.99%)	153(100.00%)
總計	34(11.00%)	98(31.00%)	132(41.00%)	55(17.00%)	319(100.00%)

若進一步就評審標準分析，表 4-5-8 顯示「困難修改」(-3）評語在量化研究的拒絕稿件及大幅修改稿件中主要為設計與概念及方法與統計，但是在接受稿件及小幅修改稿件中亦同；而在非量化研究中拒絕稿件及大幅修改稿件的「困難修改」(-3）評語係以結果討論及設計與概念為主，在接受稿件及小幅修改稿件也是相同情況。再就稿件領域來看，社會科學研究在拒絕稿件最主要的評語為貢獻度相關、設計與概念，以及方法與統計，在小幅修改稿件亦以設計與概念及方法與統計，在接受稿件則都出現在結果討論；而在人文科學研究，在拒絕稿件中多出現在結果討論、設計與概念，以及貢獻度相關，在小幅修改及接受稿件亦同。

研究者認為此一結果顯示評審者的評語與出版建議之間確實有不一致的情況，因為相同的評語、相同的評分，卻有截然不同的出版建議。本研究建議評審者在提出「困難修改」(-3）評語時，應該考慮給予拒絕或大幅修改之出版建議，以免造成期刊編輯的困擾；而期刊編輯在遇到類似情況時，也應與評審者釐清其評語之目的，以及確認其出版建議之立場。

doi:10.6342/NTU201602255

表 4-5-8 修正困難評語(-3)之評審標準與出版建議分析－依稿件研究類型(n=147)

評審標準	評審者出版建議（%）				
	拒絕	大幅修改	小幅修改	接受	總計
稿件研究類型					
量化研究					
貢獻度相關	18.75	14.29	8.33	0.00	14.04
寫作與呈現	0.00	0.00	0.00	0.00	0.00
設計與概念	25.00	28.57	41.67	100.00	31.58
方法與統計	25.00	28.57	33.33	0.00	28.07
結果討論	12.50	14.29	16.67	0.00	14.04
文獻分析	18.75	14.29	0.00	0.00	12.28
非量化研究					
貢獻度相關	20.51	10.71	27.78	0.00	17.78
寫作與呈現	5.13	3.57	0.00	0.00	3.23
設計與概念	28.21	32.14	27.78	40.00	30.00
方法與統計	0.00	0.00	0.00	0.00	0.00
結果討論	35.90	35.71	38.89	60.00	37.78
文獻分析	10.26	17.86	5.56	0.00	11.11
稿件領域					
社會科學研究					
貢獻度相關	21.43	12.50	0.00	0.00	12.73
寫作與呈現	0.00	3.13	0.00	0.00	1.82
設計與概念	21.43	31.25	50.00	0.00	30.91
方法與統計	21.43	18.75	25.00	0.00	20.00
結果討論	14.29	21.88	25.00	100.00	21.82
文獻分析	21.43	12.50	0.00	0.00	12.73
人文科學研究					
貢獻度相關	19.51	12.50	27.27	0.00	18.48
寫作與呈現	4.88	0.00	0.00	0.00	2.17
設計與概念	29.27	29.17	27.27	60.00	30.43
方法與統計	2.44	8.33	9.09	0.00	5.43
結果討論	34.15	29.17	31.82	40.00	32.61
文獻分析	9.76	20.83	4.55	0.00	10.87

另外，表 4-5-9 顯示在未出現「困難修改」(-3）評語的稿件中亦有 10%為拒絕稿件，在量化研究是以貢獻度相關、方法與統計，以及寫作與呈現的評語為主；

doi:10.6342/NTU201602255

在非量化研究則主要為寫作與呈現，68.75%。在社會科學研究中，以貢獻度相關、寫作與呈現最多，其次為方法與統計及結果討論；而在人文科學中，以寫作與呈現為最多占 64.71%。

綜合而言，研究者以為評審者若沒有提出「困難修改」(-3) 評語則不應該直接拒絕，若是基於潛在可修改之問題如寫作與呈現或文獻分析等，則更應給予作者修改機會；而期刊編輯也應提醒評審者並再次確認其出版建議之立場。

doi:10.6342/NTU201602255

表 4-5-9 未出現修正困難評語（-3）之評審標準與出版建議分析－依稿件研究類型（n=319）

評審標準	評審者出版建議（%）				
	拒絕	大幅修改	小幅修改	接受	總計
稿件研究類型					
量化研究					
貢獻度相關	22.22	17.65	20.55	36.36	20.92
寫作與呈現	22.22	43.14	30.14	9.09	32.03
設計與概念	5.56	1.96	6.85	0.00	4.58
方法與統計	22.22	9.80	15.07	27.27	15.03
結果討論	16.67	15.69	16.44	27.27	16.99
文獻分析	11.11	11.76	10.96	0.00	10.46
非量化研究					
貢獻度相關	12.50	31.91	18.64	20.45	22.29
寫作與呈現	68.75	34.04	33.90	34.09	37.35
設計與概念	6.25	6.38	11.86	13.64	10.24
方法與統計	0.00	0.00	0.00	0.00	0.00
結果討論	0.00	14.89	25.42	31.82	21.69
文獻分析	12.50	12.77	10.17	0.00	8.43
稿件領域					
社會科學研究					
貢獻度相關	23.53	25.00	20.00	33.33	22.89
寫作與呈現	23.53	38.16	32.86	33.33	34.34
設計與概念	5.88	3.95	5.71	33.33	5.42
方法與統計	17.65	6.58	12.86	0.00	10.24
結果討論	17.65	13.16	17.14	0.00	15.06
文獻分析	11.76	13.16	11.43	0.00	12.05
人文科學研究					
貢獻度相關	11.76	22.73	19.35	23.08	20.26
寫作與呈現	64.71	40.91	30.65	28.85	35.29
設計與概念	5.88	4.55	12.90	9.62	9.80
方法與統計	5.88	0.00	3.23	5.77	3.92
結果討論	0.00	22.73	24.19	32.69	24.18
文獻分析	11.76	9.09	9.68	0.00	6.54

doi:10.6342/NTU201602255

第六節　綜合討論

期刊同儕審查的理性檢驗自 1980 年代前後開始受到重視，不過直到今天利用同儕審查的評審檔案如評審報告或編輯拒絕信函等進行的研究仍然不多（Bornmann, 2011b; Chubin & Hackett, 1990; Weller, 2002）。本研究蒐集 3 種台灣出版之社會暨人文科學的 A 級或 B 級期刊、1 年或 2 年的同儕審查稿件共 48 篇、合計 103 份評審報告，利用內容分析法探討評審者使用的實際標準、評審信度，以及評審的公平性；此外本研究亦以管理學的觀點提出評審的課責性議題，希望透過評審者的自覺，提升評審評語與出版建議的合理性，也讓期刊編輯更加瞭解評審評語與出版建議之間應具有之關聯性，主要研究結論如下：

一、評審標準因著稿件研究類型與出版建議而有差異

本研究發現量化研究與非量化研究所強調的評審標準確有不同，除了寫作與呈現評語外，量化研究比較重視貢獻度相關及方法與統計，非量化研究則特別強調結果討論。此外在不同的出版建議，評審者的評語亦呈現不同重點，例如在量化研究中，拒絕稿件最重要的負面評語首為方法與統計，其次為貢獻度相關，接受稿件的正面評語則以貢獻度相關最為優先；而在非量化研究中，拒絕稿件的負面評語主要為結果討論，其次為設計與概念，至於接受稿件的正面評語則以結果討論及貢獻度相關最為重要。

二、接受稿件的評審信度高於拒絕稿件

本研究將出版建議分為3種模式討論[4]，無論是模型A、模型B或模型C，拒絕稿件的Cohen's Kappa值均低（0.00至0.32），接受稿件的Kappa值較高（1.00至0.42）；至於大幅修改稿件及小幅修改稿件之評審出版建議信度亦在0.38至0.42之間。研究

[4]出版建議之 3 種分析模式：1. 評模式 A（4 類）：拒絕、大幅修改、小幅修改，以及拒絕；2. 模式 B（3 類）：拒絕、大幅修改，以及接受（含小幅修改）；3. 模式 C（2 類）：拒絕（含大幅修改）及接受（含小幅修改）。

doi:10.6342/NTU201602255

者推論我國社會暨人文科學的評審者，對於拒絕稿件的寬嚴程度差異較大，但是在接受、大幅修改，以及小幅修改稿件的的共識程度則較高。然而許多研究指出期刊評審者對於拒絕稿件的一致性較高（Cicchetti, 1991, 1997; Daniel, 1993; Ingelfinger, 1974），而且是接受稿件的2倍（Weller, 2001）。

三、非量化研究及人文科學領域的評審信度高於量化研究

就出版建議的評審信度來看，本研究在量化研究的 Cohen's Kappa 值在 3 種出版建議模式中都小於 0，非量化研究的 Cohen's Kappa 值則較高，在 0.23 至 0.46 之間，而根據後設分析文獻，期刊同儕審查出版建議的 Cohen's Kappa 平均值為 0.17。至於評審標準的評審信度，量化研究的 Cohen's Kappa 值亦低於非量化研究，各分別為 0.22 及 0.33。總之儘管許多學者認為量化研究有相同的評審標準，評審原則較為一致，不過在本研究中卻沒有呈現相對較高的評審信度。

四、評審者的審查風格因著稿件研究類型與稿件領域而有不同

本研究以非科學品質評語、情緒性用語，以及透露出版建議做為潛在公平性類目，發現所有評審者均只針對科學品質進行審查，未有提出非科學品質評語。不過研究者亦發現評審者使用情緒性用語及透露出版建議的情況，比例雖然不高，分別為 5.83%及 10.68%，但是有 80%以上出現在非量化研究及人文科學領域，亦即非量化研究及人文科學研究之評審者有比較特殊的審查文化。

五、評審公平性的因果推論不易，某些可透過編訂評審指南避免

本研究的評審者未有提出非科學品質之評論，不過本研究亦發現評審者使用情緒性用語之情況，Spencer 等人（1986）以心理學期刊的評審報告進行研究發現，使用情緒性或不確定評論超過 25%，推論可能受到評審者個人偏見的影響，本研究之情緒性用語之比例僅有 5.83%，雖大多為拒絕稿件，但是難以論斷為評審者之偏見。另有 10.68%的評審者在評語中透露出版建議，或許存有潛在公平性議題，

doi:10.6342/NTU201602255

但亦必需更嚴謹的推論其成因。因此本研究認為某些公平性議題，可以透過編訂評審指南，即可實質改進此一情況。

六、評審評語與出版建議存在不一致的情況

為分析評審評語與出版建議的一致性，本研究將正面評語依「具體性」、負面評語依「可修改性」分別討論，亦即評審者在建議接受稿件時，應具體指出稿件的優點，而在拒絕稿件時亦應提出改正困難之批評。就本研究的正面評語之「具體性」分析發現，量化研究的「很具體」(+3) 評語中有 75%在接受稿件，非量化研究則僅 57.14%；而且量化研究的「不具體」評語屬於接受稿件者為 0.00，而非量化研究有 25.00%，推論非量化研究比較可能出現正面評語無法支持其接受稿件之出版建議。

而就負面「可修改性」評語來看，發現確實存在評審者的評語與出版建議不一致的情況，例如量化研究與非量化研究在接受稿件及小幅修改稿件的負面「困難修改」(-3) 評語都占 1/4 左右，而且其主要評語亦與拒絕及大幅修改稿相同；此外在沒有出現「困難修改」(-3) 評語的評審報告中，亦有 10%為拒絕稿件。據此研究者認為評審者若提出「困難修改」(-3) 之評語，則應給予拒絕出版之建議，以免造成編輯的困擾；而若未提出「困難修改」(-3) 之評語者，則應給予作者再次修改的機會，而不宜直接拒絕

七、評審者對拒絕稿件的寬嚴程度依稿件研究類型及稿件領域有別

本研究發現量化研究與非量化研究的評審者在拒絕稿件時有不同的寬嚴程度，在量化研究中的「困難修改」(-3) 評語多出現在大幅修改稿件 (49.12%)，非量化研究則為拒絕稿件 (43.33%)，亦即量化研究的評審者雖然提出「困難修改」(-3) 的意見，卻仍多給予作者大幅修改之出版建議，讓作者獲得再次審查的機會；至於非量化研究的評審者在給予「困難修改」(-3) 的評語後，則傾向直接拒絕，此一情況在稿件領域更為明顯，社會科學領域研究中的「困難修改」(-3) 評語多出

doi:10.6342/NTU201602255

現在大幅修改稿件（58.18%），非量化研究則為拒絕稿件（44.57%）。研究者推論其原因之一可能受到期刊拒稿率的影響，在高拒稿率期刊，評審者需要選出最卓越的稿件，至於不適合者則直接拒絕；而在低拒稿率的期刊，囿於稿件選擇有限，評審者比較不會直接拒絕稿件，此也反應期刊稿件審查除了基於科學品質外，亦受到審查過程情境的影響（Bornmann, 2010）。

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

第五章　結論與建議

同儕審查是學術品質評鑑與資源分配的主要機制之一，在期刊文獻出版、獎助資源分配，以及大學教職的聘用與升遷上都扮演著關鍵角色，許多國家並將之應用於科技政策制定及教研機構評鑑等業務。近半個多世紀以來，同儕審查的效用、效率、信度，以及公平性受到多方質疑，有關同儕審查的研究也因而日益受到重視，研究的廣度與深度亦逐漸增加，而成為一個獨立且益形重要的學門。以下依本研究之研究目的統整歸納研究結果並提出相關建議，全章分為研究結論、研究建議、研究貢獻，以及未來研究方向等節說明。

第一節　研究結論

本研究目的除了回顧同儕審查機制的起源、研究現況與發展外，並以 3 種我國發行之社會暨人文科學期刊的同儕審查評審報告進行實徵研究，探討我國期刊同儕審查的評審標準、信度與公平性，另亦兼及評審的課責性議題。基於前述目的，本節研究結論分為三部分，首先綜合討論同儕審查機制之現況與發展，其次彙析同儕審查的研究現況與困境，最後提出對於我國社會暨人文科學期刊同儕審查研究之省思。

一、同儕審查機制之現況與發展

（一）同儕審查評審機制逐漸開放與透明

同儕審查機制的理性檢驗在 1980 年代前後逐漸增加，不過因為第一手評審檔案資料取得不易，研究結果有其侷限。直到 20 世紀末期，西方國家的公共獎助機構為提升同儕審查的品質與效能，逐漸願意將評審檔案提供學者進行研究，有些機構甚至允許學者進行評審者訪談或是觀察評審委員會的議事決策過程，使得獎

doi:10.6342/NTU201602255

助同儕審查機制逐漸邁向開放與透明之途。至於期刊同儕審查主事機構對於一手評審報告的開放迄今雖然仍多保留，但是卻有許多創新作法，例如預印同儕審查、出版後同儕審查，以及公開同儕審查等，這些改革都是朝向開放與透明的審查機制前進。

（二）同儕審查的國際交流持續進行

同儕審查主事機構為了提升評審作業的品質與公信力，積極進行國際交流，彼此分享經驗。目前期刊及獎助同儕審查都已有跨國性的合作成果，編訂同儕審查作業之參考指南；而大學教職聘用／升遷同儕審查亦有跨國研究計畫，提出有關個人學術成就之基礎評審指標。此外美國等近 50 餘國的獎助機構復於 2012 年設立全球研究委員會（GRC），除了提出科學同儕審查之原則聲明外，並積極推動跨洲際之間的多邊研究與合作。該委員會每年在不同國家舉辦年會，過去 3 年的主辦國分別為德國、中國及日本，本（2016）年年會已在印度舉行，討論主題為女性在研究領域的平等地位及跨領域研究。

（三）同儕審查之評審品質與效能提升的作法多元

除了對於同儕審查作業的理性檢驗外，如何透過良好的作業管理機制以確保評審品質並提升效能，也受到學術界及同儕審查主事機構的重視，相關作法甚為多元。以期刊同儕審查為例，有試圖分析優良評審者的特質；有編訂評審指南說明評審者的權責；有加強評審者的審查訓練；有建置評審者評鑑系統，以及採用實驗方式評鑑評審者表現等。此外許多獎助機構亦認知到同儕審查作業具有保守主義傾向且不利年輕科學家等，而採取了各種不同的補強作法，例如聘請非專業人士進入評審團隊，強化獎助計畫管理者的決策權限，以及針對某一群體或特定目標設置專屬獎助項目等　。

doi:10.6342/NTU201602255

（四）同儕審查機制之展望

1、**建立持續監督、檢驗與改進的機制：**就社會學的觀點，同儕審查透過一個持續性、去中心性，以及社會分擔的錯誤嘗誤過程，讓創新研究得以進行；至於同儕審查的正當性則是奠基於學術社群成員之間的彼此信賴與誠信。不過學術領域越來越專精且複雜，研究人口也愈來愈多，在學術資源未見大幅成長的背景下，競爭將日趨激烈，同儕審查作業勢必面對外界更多的質疑與挑戰，因此建立一個持續監督、檢驗與改進的機制，應是學術界的共同努力方向。

2、**同儕審查與書目計量的競合關係：**質性的同儕審查與量化的書目計量是當前學術領域評鑑的兩大主軸，兩種方式各有其優缺點，而廣泛客觀的書目計量指標應有助於提升同儕審查的合理性與透明性。目前同儕審查與書目計量的整合評鑑方式，已應用在機構層級的學術評鑑，例如國家級高教機構評鑑等；而有些研究利用引用數據作為期刊同儕審查之效度指標，也是另一種形式的整合。但是同儕審查與書目計量共同運用在個人層級的評鑑時就引發較多爭議，尤其是在獎助計畫及大學教職聘用／升遷同儕審查，包括如何選擇適當的書目計量指標、如何兼顧領域之間的差異性，以及如何分配兩者的權重等，均仍有許多討論空間。

二、同儕審查的研究現況與發展

（一）**評審標準的研究依然受到重視：**評審標準一向是同儕審查研究的核心，以往大多針對審查過程中的利益關係者進行所謂規範標準的意見調查，只有少數利用一手評審者的審查報告進行實際標準研究。20 世紀末，歐美等國家的政府獎助機構陸續開放評審檔案提供研究，進入了評審者實際標準的探討，頗有助於瞭解獎助同儕審查的內部運作機制；而期刊同儕審查研究也逐漸重視評審者使用之實際標準與出版建議之間的一致性。此外長期以

doi:10.6342/NTU201602255

來大學教職聘用／升遷同儕審查重視研究並輕忽教學之作法，也已引起學術界人士的關切，希望訂定個人學術表現評鑑之規範標準，歐盟第 7 架構支持的 ACUMEN 計畫即為一例。

（二）評審信度過低的因果研究有待加強：學術界對於同儕審查評審信度過低的看法不一，有的認為是科學審查的常態，有的因而質疑評審作業的公平性；此外有關評審之間低信度的成因研究，迄今仍未有系統性結論。總之同儕審查的低信度已是一個事實，分析評審者的審查報告或可找出其成因；不過影響評審信度的面向甚為廣泛，在分析時亦需考量評審者的特殊習性與寬嚴態度，以及個別期刊特有之審查程序或拒稿率等，此外強化研究效度與合理的樣本數亦有助於提升因果推論之可信度。

（三）公平性研究的方法論受到質疑：目前同儕審查研究已經證實存有評審不公的情況，發現各種不同類型的評審者偏見，但是此類研究經常受到批評，包括方法論不夠嚴謹、因果推論薄弱，而且各篇研究結果的異質性甚大難以通則化。有學者認為利用實驗性研究或可補強公平性研究方法論之不足，但是卻有易觸學術倫理紅線之憂。近年來已有學者利用後設分析法進行公平性的通則性推論，但是研究結果亦有差異，需要更多實徵研究加以驗證，而大規模一手評審檔案的分析仍為公平性研究的可行之路。

（四）同儕審查的效用亟待證實：同儕審查作為學術界有限資源的重要分配機制之一，但是其效用卻迄未證實，除了因為缺乏效用評定的指標外，研究操作不易也是其主要限制。一般來說，同儕審查的效用研究除了針對同儕審查中的勝利者進行分析外，亦需比較失敗者的後續發展。基此，期刊同儕審查的效用研究相對較為容易，因為許多稿件退稿後在其他期刊登載的比例甚高。但是落選的獎助計畫則不然，許多計畫大多胎死腹中，無法進行評審效用的比對與檢驗。至於未受聘大學教師之追蹤研究亦甚為不易，影響其成就發展的變數也更加難以預期。

doi:10.6342/NTU201602255

三、我國社會暨人文科學學期刊同儕審查研究之省思

（一）評審報告的內容分析呈現評審者實際使用之核心標準

理論上研究評審標準最直接的方式是利用第一手的評審者審查報告，本研究透過評審報告的內容分析，呈現評審者在審查時最常使用的核心標準。另並透過稿件研究類型的討論發現，除了寫作與呈現評語外，量化研究比較重視貢獻度相關及方法與統計；非量化研究則特別強調結果討論，至於有關方法與統計的意見則闕如。另外若就稿件領域分析時發現，社會科學與人文科學研究之核心標準大多呼應稿件研究類型的結果，研究者推論在學門層級的討論可能會呈現較明顯之差異。

（二）出版建議分析可細部解讀拒絕稿件或接受稿件之主要理由

本研究利用內容分析法將評審報告內容分類及區分不同面向之正負等級，另並透過出版建議進行討論，探討評審者在接受稿件時最重要的正面評語，或是在拒絕稿件時之主要負面意見。例如在量化研究中，接受稿件的正面評語以貢獻度相關最為優先，而拒絕稿件最重要的負面評語首為方法與統計，其次為貢獻度相關。而在非量化研究中，接受稿件的正面評語以結果討論及貢獻度相關最為重要，至於拒絕稿件的負面評語主要為結果討論，其次為設計與概念。

（三）稿件研究類型及領域分析呈現非量化研究及人文科學研究的評審者特性

本研究在進行稿件研究類型分析時發現，非量化研究除了在評審標準與量化研究不同外，另有 3 項評審特色，其一、無論在評審者的評審標準或出版建議，評審信度都比較高；其二、評審者若提出「修改困難」（-3）之評語，大多直接建議為拒絕稿件，不讓作者有再次審查之機會；其三、評審者較常使用情緒性用語及向作者透露出版建議。而在比較稿件領域時，人文科學研究也有類似的特色，研究者推論可能是因為人文科學領域的非量化研究較多的緣故。

doi:10.6342/NTU201602255

（四）公平性議題需審慎推論因果

本研究並未發現評審者使用非科學品質評語，不過確實有評審者使用情緒用語及向作者透露出版建議之情況。儘管有學者認為情緒用語可能受到評審者個人偏見的影響（Spencer et al., 1986），本研究認為情緒性用語或許可作為潛在公平性議題，但是必需要進行大樣本及長時期的分析，並利用廣泛變項審慎推論成因。至於向作者透露出版建議一事，研究者認為無論是否影響評審的公平性都應該明白禁止，因為臺灣是個小型科研國家，評審者大多可以辨識作者之身分，出版建議訊息的透露，容易引起不必要的聯想。

（五）評審評語與出版建議的一致性應進行系統討論

為探討評審評語與出版建議的一致性，本研究將正面評語以「具體性」、負面評語以「可修改性」進行正負 3 級距的評分發現，評審者的評語與出版建議確實出現不一致的情況，因而提出評審的課責性議題。例如評審者若提出「困難修改」（-3）的評語，就不宜提出接受稿件的建議；而若沒有提出「困難修改」(-3）的意見，亦不宜直接拒絕稿件。此外期刊編輯在面對評審評語與出版建議不一致時，亦應聯繫評審者以確認其立場。研究者認為有關評審的課責性議題，若能有系統討論，對於提升評審作業品質應有助益。

（六）評審者行為理論研究有其必要性

本研究發現量化研究與非量化研究的評審者在拒絕稿件時有不同的寬嚴態度，量化研究的評審者儘管提出「困難修改」(-3）之評語，仍多給予大幅修改的出版建議，讓稿件獲得再次審查之機會；而非量化研究的評審者在相同情況則傾向直接拒絕。研究者推論其原因之一可能受到期刊稿件拒稿率的影響，在高拒稿率期刊，評審者需要選出最卓越的稿件，對於不適合者則直接拒絕；而在低拒稿率的期刊，囿於投稿數量，評審者較少直接拒絕稿件。總之，影響評審者行為的因素很多，期刊評審者在審查時的思辨與決策過程的理論研究，應有助於評審經驗的

doi:10.6342/NTU201602255

傳承與學習，以及提升同儕審查評審機制的運作效能。

（七）我國人文科學研究有其特殊的撰稿方式

本研究為將研究結果與相關文獻進行比較，以呈現我國期刊同儕審查作業之特色，評審報告評語之主題編碼係採自 Bornmann、Nast 等人（2008）的後設分析文獻所提出之期刊同儕審查評審標準分類架構。不過研究者在利用該分類架構編碼時發現，我國人文科學研究有其獨特之撰稿方式，稿件大部分內容係針對某個概念進行貫通古今之辯證，與科學領域量化研究的撰寫方式頗有差異，亦與非量化研究有別，使得研究者在歸類人文科學研究之評語時，常有捉襟見肘、削足適履之感。

第二節　研究建議

同儕審查為學術領域有限資源的分配機制，近半個多世紀以來，西方學者已將同儕審查視為知識研究的主題，為其建構科學研究的價值與方向；反觀我國政府與學術界對於同儕審查仍多流於私下批評，研究與討論均嫌不足。本研究首先針對我國同儕審查機制與研究的整體發展提出建議，其次則針對我國社會暨人文科學期刊同儕審查之研究結果提出看法。

一、我國同儕審查機制與研究的整體發展建議

（一）營造開放透明的同儕審查環境，提升國家科研競爭力

西方學者對於同儕審查的理性檢驗超過半個世紀，目前已有愈來愈多利用第一手評審檔案進行的研究，在不久的將來一個開放且透明的同儕審查機制可期。反觀我國關於同儕審查的研究僅有少數論述式文獻，學術界並未將之視為知識研究的主體，我國各類同儕審查主事機構亦未有系統地進行評審程序的自我檢驗。目前我國已於 2005 年施行政府資訊公開法，保障民眾對公務資訊知的權力，同儕

doi:10.6342/NTU201602255

審查的一手評審檔案在兼顧個人資訊保護法的前題下，政府相關機構應主動開放學界進行研究，以強化我國同儕審查之評審品質、促進學術資源分配效率，以及健全學術研究環境，進而提升國家科研的競爭力。

（二）政府結合學術界鼓勵進行同儕審查相關研究

整體來看，同儕審查研究大致可分為兩種類型，其一為同儕審查主事機構的自我檢驗或委託研究，其二為學術界的自發性研究。前者政府可以績效管考之作法予以強化，例如要求政府獎助機構或公立教育或研究單位，定期就其主辦之各項同儕審查作業進行評審流程與效能的自我檢驗，並以之做為機構績效考評的參考。此外為鼓勵學術界進行相關研究，政府可贊助教研機構舉辦同儕審查研討會、推動同儕審查研究計畫，以及出版同儕審查文獻或專書等。

（三）參與同儕審查之國際合作

在知識經濟的社會，科學研究是各國政府進行政治、經濟，以及社會決策時的重要參考，更在國家推動創新過程中扮演關鍵性角色。今天歐美等西方國家都體認到良好的同儕審查機制有助於提升國家的競爭力，因而積極推動同儕審查的國際合作與交流，以強化評審作業的品質與公信力。我國政府相關機構亦宜主動參與或鼓勵學術界進行同儕審查國際交流，例如加入全球研究委員會（GRC）或推廣國際醫學期刊編輯委員會（ICMJE）之學術期刊規範等，以提升我國同儕審查之作業品質。

（四）逐步發展各類同儕審查之監督與檢驗機制

同儕審查已是各項學術活動的主要仲裁者，目前許多世界著名之同儕審查期刊的稿件接受率只有個位數字，各國政府與民間獎助機構之獲獎率也有逐年下降的趨勢，而大學教職聘用與升遷的評審方式亦已受到公平性的批評。因此未來同儕審查作業勢必面對外界更多的質疑與挑戰，我國對於同儕審查的研究雖然仍在起步階段，但是若能針對不同類型的同儕審查，逐步發展各種監督、檢驗與改進

doi:10.6342/NTU201602255

的機制，對於我國同儕審查作業應有實質性的助益。

二、有關我國社會暨人文科學期刊同儕審查作業之建議

（一）有系統探討評審的課責性議題

同儕審查的公平性研究迄今未有通則性結論，強化評審的課責性也許是提升我國期刊同儕審查品質的可行作法之一。本研究以評審者的評語與出版建議的一致性來探討評審的課責性，希望評審者能更加重視其評語與出版建議之間的合理性；同時也讓期刊編輯瞭解評審者的評語與出版建議的關聯性，達到提升期刊同儕審查之評審品質及強化公平性的目的。未來研究若能系統性討論評審者或期刊編輯在同儕審查之權利與責任，相信有助於健全我國期刊同儕審查機制。

（二）編訂我國社會暨人文科學期刊稿件投稿及同儕審查作業規範

本研究分析我國社會暨人文科學期刊的評審報告，發現評審者的評語與出版建議不一致、使用情緒性用語以及向作者透露出版建議等情況。為提升我國社會暨人文科學之期刊同儕審查之品質，似宜仿效國際期刊編輯委員會（ICMJE）的作法，由學術界合作編訂我國社會暨人文科學期刊稿件投稿規範，內容除包括稿件撰寫標準外，並詳述稿件作者、評審者、期刊編輯及出版者在同儕審查過程中的職能與責任，以達到實質改進期刊同儕審查品質之效用。

（三）推動同儕審查評審訓練課程，整體提升同儕審查之評審品質

同儕審查的評審經驗傳承，大多為教授帶領個別研究生的師徒制，或由學者邊做邊學。近年來有些獎助機構提供同儕審查訓練課程，有些期刊編輯讓評審者互相觀摩評審報告，以提升審查品質。本研究發現我國社會暨人文科學期刊同儕審查，評審報告的總字數與評語筆數之差異甚大、評審者在評審標準與出版建議的信度均不高，而且出現評審者使用情緒性用語，以及評審評語與出版建議不一致的情況。綜此，我國政府及教育機構若能針對我國博、碩士生提供短期或正式的評審訓練課程，除教導學術研究之品質審查技巧外，並兼討論評審道德與誠信

doi:10.6342/NTU201602255

等議題，對於提升同儕審查之評審品質應有治本效用。

（四）將期刊同儕審查的自我檢驗列為社會暨人文科學期刊評比的項目之一

我國科技部為促進社會暨人文科學的研究發展及提升期刊編輯水準，於 1990 年代開始推動社會暨人文科學期刊評比，本（2016）年公布的實施方案，評審標準包括出版格式、論文格式、編輯作業，以及刊行作業四大項，申請評比之期刊需提出同儕審查的作業過程資訊，包括評審委員會組成、評審會議結果、評審者的審查報告等。研究者認為若能將期刊同儕審查作業的自我檢驗列為評審項目之一，除有助於各期刊進行同儕審查之自我品質管考外，並有助於鼓勵同儕審查研究。

第三節　研究貢獻

根據本研究之研究對象、研究方法，以及研究結果，主要貢獻如下：

一、研究對象的特色

本研究係我國首篇整體回顧同儕審查機制之起源、研究現況與發展之論文，也是首篇針對我國出版之社會暨人文科學期刊的評審報告進行之實徵研究，除了分析評審報告內容外，並透過出版建議、稿件研究類型，以及稿件領域等變項，探討我國期刊同儕審查之評審標準、評審信度、評審的公平性及課責性議題。

二、研究設計的創新

本研究認為評審者在進行稿件審查時有其權責，在建議接受稿件時，應具體指出稿件的優點；而在建議拒絕稿件時，亦應提出難以改正的批評，因此提出評審的課責性議題。本研究首創將評審者的正面評語依「具體性」、負面評語依「可修改性」進行正負面向的分級，以探討評審評語與出版建議的一致性。

doi:10.6342/NTU201602255

三、研究結果的價值

（一）探討同儕審查機制之起源、研究現況與未來發展，有助於我國學術界瞭解同儕審查做為知識研究主體的歷程與現況發展。

（二）利用我國社會暨人文科學期刊之同儕審查報告進行實徵研究，呈現我國社會暨人文期刊同儕審查之特色，包括評審實際使用之核心標準、拒絕稿件或接受稿件之主要理由、非量化研究及人文科學研究的評審特色、以及人文科學研究的特殊撰稿方式等。

（三）提出期刊同儕審查評審品質提升的具體作法：本研究提出評審的課責性議題，將正面評語依「具體性」、負面評語依「可修改性」進行分析，探討評審者評語與出版建議的一致性，可在實務上提升評審作業品質。

第四節　未來研究方向

根據本研究之研究限制、研究設計，以及研究結果，提出未來相關研究建議，以做為日後進一步研究之參考：

一、深入介紹各類同儕審查之研究現況與發展

本研究目的之一係以期刊同儕審查、獎助同儕審查，以及大學教職聘用／升遷同儕審查為例整體回顧同儕審查機制之起源、研究現況與發展，但是對於個別同儕審查之當前重要議題未能詳細介紹，例如有關期刊同儕審查的創新模式、獎助同儕審查之社會影響指標或創新指標之評審作業，個人學術成就之評審標準架構，以及同儕審查的誠信問題等都是值得深入分析的主題，此類專題式的回顧文獻對推動我國同儕審查研究甚有助益。

doi:10.6342/NTU201602255

二、期刊同儕審查之評審課責性需要更多面向的討論

本研究首創將評審者的正面評語依「具體性」、負面評語依「可修改性」進行討論，發現確實存在評審者評語與出版建議不一致的情況，而提出強化評審的課責性議題。鑒於期刊編輯與評審者都是學術傳播的守門者，其權利需要被監督與檢驗，未來若能系統性討論評審者或期刊編輯的課責性議題，相信有助於健全期刊同儕審查機制。

三、單種期刊、大樣本之研究有其必要性

本研究採便利抽樣，以國內出版之 3 種社會暨人文科學期刊為研究對象，針對評審者評語及出版建議進行分析，但是為保護資料來源，無法針對個別期刊的特色進行討論。未來研究建議以單種期刊、長時期的大樣本進行研究，如此可以利用期刊本身之特質如拒稿率、評審者遴選、評審委員會組成、評審作業程序、評審評分系統等進行檢驗；此外若能同時討論評審者及作者背景，應有助於強化期刊同儕審查公平性及信度議題的因果分析。

四、評審者行為理論研究，有助評審經驗傳承

本研究發現評審者的審查寬嚴有別，而且評審者評語與出版建議也存在不一致的情況，例如有的評審者提出「困難修改」(-3) 之意見，卻給予接受稿件的建議；而有的評審者未提出「困難修改」(-3) 之評語，卻建議接受，其成因為何，值得進一步探討。總之影響評審者出版建議的因素很多，需在廣泛基礎上進行研究，而且評審者在做決定前之內省以及直覺，是否在評審報告中顯現仍未可知，透過評審者行為理論研究應可有所釐清，也可理解期刊的審查規範是如何的流傳、學習與執行。

五、編訂我國人文科學研究之撰稿規範與評審標準

本研究為了與目前相關研究進行比較，使用 Bornmann、Nast 等人（2008）之

doi:10.6342/NTU201602255

後設分析文獻提出之期刊同儕審查評審標準分類架構，惟在進行人文科學研究分類時，常有捉襟見肘之感。鑒於我國歷史悠久，許多人文科學文獻在研究取材及撰稿方式上均有其特殊性，未來若能針對我國人文科學研究編訂稿件撰寫規範及同儕審查評審標準，將有助於我國人文科學領域之研究發展，以及提升同儕審查機制之品質。

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

參考文獻

卯靜儒（2013）。學術期刊的同儕審查為哪樁？維持品質？鼓勵創新？*臺灣教育評論月刊，2*(9)，13-16。

成群豪（2006，11月）。知識產業化對大學治理之啟示與探究。*2006第二屆兩岸高等教育論壇：高等教育質量、辦學模式及其發展策略學術研討會論文集*（頁143-161）。廣州：華南理工大學。

林娟娟（1997）。學術期刊之同儕審查。*大學圖書館，1*(3)，127-140。

邱炯友（2003）。學術電子期刊同儕評閱之探析。*教育資料與圖書館學，40*(3)，309-323。

陸偉明（2009）。同儕審查制度。*人文社會科學簡訊，10*(4)，117-123。

黃宗貴（2010a）。大學治理機制之相關理論與運作模式。*網路社會學通訊期刊，88*。檢自http://society.nhu.edu.tw/e-j/88/88-24.htm

黃宗貴（2010b）。*歐美大學內部治理機制之研究*（未出版之碩士論文）。國立嘉義大學教育行政與政策發展研究所，嘉義市。

黃毅志、曾世杰（2008）。教育學術期刊高退稿率的編審制度、惡質評審與評審倫理。*台東大學教育學報，19*(2)，183-196。

楊開煌（1998）。中共研究中的內容分析法及其爭議與反省。*東亞季刊，20*，2-28頁。

Abbott, A. (2008, June). *Publication and the future of knowledge*. Paper presented at the Association of American University Presses, Montreal, Canada. Retrieved from http://home.uchicago.edu/~aabbott/Papers/aaup.pdf

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., Durand-Zaleski, I., & Alberti, C. (2012). Peer review of grant applications: Criteria used and qualitative study of reviewer practices. *PLoS ONE, 7*(9), e46054. doi: 10.1371/journal.pone.0046054

Abdoul, H., Perrey, C., Tubach, F., Amiel, P., Durand-Zaleski, I., & Alberti, C. (2012). Non-financial conflicts of interest in academic grant evaluation: A qualitative study of multiple stakeholders in France. *PLoS ONE, 7*(4), e35247. doi: 10.1371/journal.pone.0035247

Abramo, G., & D'Angelo, C. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics, 87*(3), 499-514. doi: 10.1007/s11192-011-0352-7

doi:10.6342/NTU201602255

ACUMEN Consortium. (2014). *Guidelines for good evaluation practice with the ACUMEN portfolio*. Retrieved from http://research-acumen.eu/wp-content/uploads/D6.14-Good-Evaluation-Practices.pdf

Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing peer review. *Science, 321*(5885), 15.

Altman, L. K. (1996). The Ingelfinger rule, embargoes, and journal peer review—Part 1. *The Lancet, 347*(9012), 1382-1386. doi: 10.1016/S0140-6736(96)91016-8

American Association for the Advancement of Science. (2010). *Congressional R&D earmarks by agency and program.* Retrieved from http://www.aaas.org/sites/default/files/migrate/uploads/earm10c.pdf

American Association of University Professors. (1915). *AAUP's 1915 declaration of principles.* Retrieved from http://aaup.org.uiowa.edu/files/aaup.org.uiowa.edu/files/Gen_Dec_Princ.pdf

American Association of University Professors. (1940). *1940 statement of principles on academic freedom and tenure.* Retrieved from http://www.aaup.org/file/1940%20Statement.pdf

American Association of University Professors. (1966). *Statement on government of colleges and universities.* Retrieved from http://www.aaup.org/report/statement-government-colleges-and-universities

Andersen, D. L. (Ed.). (2003). *Digital scholarship in the tenure, promotion, and review process*. New York, NY: M. E. Sharpe.

Armstrong, J. S. (1996). We need to rethink the editorial role of peer reviewers. *Chronicle of Higher Education, 43*(9), B3-B4.

Baxt, W. G., Waeckerle, J. F., Berlin, J. A., & Callaham, M. L. (1998). Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of Emergency Medicine, 32*(3), 310-317. doi: 10.1016/S0196-0644(98)70006-X

Bailar, J. C. (1991). Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behavioral and Brain Sciences, 14*(1), 137-138. doi: 10.1017/S0140525X00065705

Bakanic, V., McPhail, C., & Simon, R. J. (1989). Mixed messages: Referees' comments

doi:10.6342/NTU201602255

on the manuscripts they review. *Sociological Quarterly, 30*(4), 639-654. doi: 10.1111/j.1533-8525.1989.tb01540.x

Becher, T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines* (2nd ed.). Buckingham, England: Open University Press.

Bedeian, A. G. (2004). Peer review and the social construction of knowledge in the management discipline. *Academy of Management Learning and Education, 3*(2), 198-216. doi: 10.5465/AMLE.2004.13500489

Bell, S., Shaw, B., & Boaz, A. (2011). Real-world approaches to assessing the impact of environmental research on policy. *Research Evaluation, 20*(3), 227-237. doi: 10.3152/095820211X13118583635792

Berelson, B. (1952). *Content analysis in communication research*. Glencoe, IL: The Free Press.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44*(2), 451-466.

Bertout, C., & Schneider, P. (2004). Editorship and peer-review at A&A. *Astronomy and Astrophysics, 420*(3), E1. doi: 10.1051/0004-6361:20040182

Bexley, E., James, R., & Arkoudis, S. (2011). *The Australian academic profession in transition.* Retrieved from University of Melbourne, Centre for the Study of Higher Education website: http://careers.unimelb.edu.au/__data/assets/pdf_file/0003/723315/The_Academic_Profession_in_Transition_Sept2011.pdf

Beyer, J. M., Chanove, R. G., & Fox, W. B. (1995). The review process and the fates of manuscripts submitted to AMJ. *Academy of Management Journal, 38*(5), 1219-1260.

Bhattacharya, A. (2012). Science funding: Duel to the death. *Nature, 488*(7409), 20-22. doi: 10.1038/488020a

Biagioli, M. (2002). From book censorship to academic peer review. *Emergences: Journal for the Study of Media & Composite Cultures, 12*(1), 11-45. doi: 10.1080/1045722022000003435

Bloch, C., Graversen, E. K., & Pedersen, H. S. (2014). Competitive research grants and

doi:10.6342/NTU201602255

their impact on career performance. *Minerva, 52*(1), 77-96. doi: 10.1007/s11024-014-9247-0

Boden, M., Ash, E., Edge, D., Reece, C., Skehel, J., & Williams, P. (1990). *Peer review: A report to the advisory board for the research councils from the working group on peer review*. Retrieved from Advisory Board for the Research Councils website: http://webarchive.nationalarchives.gov.uk/20160217110318/http://mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC003951

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: MIT Press.Bornmann, L. (2010). Does the journal peer review select the "best" from the work submitted? The state of empirical research. *IETE Technical Review, 27*(2), 93-96. doi: 10.4103/0256-4602.60162

Bornmann, L. (2011a). Peer review and bibliometric: Potentials and problem. In J. C. Shin, R. K. Toutkoushian, & U. Teichler (Eds.), *University rankings: Theoretical basis, methodology and impacts on global higher education* (pp. 145-164). Berlin, Germany: Springer. doi: 10.1007/978-94-007-1116-7_8

Bornmann, L. (2011b). Scientific peer review. *Annual Review of Information Science and Technology, 45*(1), 197-245. doi: 10.1002/aris.2011.1440450112

Bornmann, L. (2012). Measuring the societal impact of research. *EMBO Reports, 13*(8), 673-676. doi: 10.1038/embor.2012.99

Bornmann, L. (2013a). Evaluations by peer review in science. *Springer Science Reviews, 2013*(1), 1-4. doi: 10.1007/s40362-012-0002-3

Bornmann, L. (2013b). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology, 64*(2), 217-233. doi: 10.1002/asi.22803

Bornmann, L., & Daniel, H.-D. (2004). Reliability, fairness and predictive validity of committee peer review: Evaluation of the selection of post-graduate fellowship holders by the Boehringer Ingelheim Fonds. *Futura, 19*, 7-19.

Bornmann, L., & Daniel, H.-D. (2005). Criteria used by a peer review committee for selection of research fellows: A Boolean probit analysis. *International Journal of Selection and Assessment, 13*(4), 296-303. doi: 10.1111/j.1468-2389.2005.00326.x

Bornmann, L., & Daniel, H.-D. (2006). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80. doi: 10.1108/00220410810844150

doi:10.6342/NTU201602255

Bornmann, L., & Daniel, H.-D. (2008). The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at *AngewandteChemie*. *AngewandteChemie International Edition, 47*(38), 7173-7178. doi: 10.1002/anie.200800513

Bornmann, L., & Daniel, H.-D. (2010). Reliability of reviewers' ratings when using public peer review: A case study. *Learned Publishing, 23*(2), 124-131. doi: 10.1087/20100207

Bornmann, L., Herich, H., Joos, H., & Daniel, H.-D. (2012). In public peer review of submitted manuscripts, how do reviewer comments differ from comments written by interested members of the scientific community? A content analysis of comments written for *Atmospheric Chemistry and Physics*. *Scientometrics, 93*(3), 915-929.

Bornmann, L., Leydesdorff, L., & van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics, 4*(3), 211-220. doi: 10.1016/j.joi.2009.10.004

Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics, 98*(1), 487-509. doi: 10.1007/s11192-013-1161-y

Bornmann, L., Marx, W., Schier, H., Thor, A., & Daniel, H.-D. (2010). From black box to white box at open access journals: Predictive validity of manuscript reviewing and editorial decisions at *Atmospheric Chemistry and Physics*. *Research Evaluation, 19*(2), 105-118. doi: 10.3152/095820210X510089

Bornmann, L., Mutz, R., & Daniel, H. (2007). Gender differences in peer reviews of grant applications: A meta-analysis. *Journal of Informetrics, 1*, 226-238. doi: 10.1016/j.joi.2007.03.001

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Latent Markov modeling applied to grant peer review. *Journal of Informetrics, 2*(3), 217-228. doi: 10.1016/j.joi.2008.05.003

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). *A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants*. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0014331

doi:10.6342/NTU201602255

Bornmann, L., Nast, I., & Daniel, H.-D. (2008). Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication. *Scientometrics, 77*(3), 415-432. doi: 10.1007/s11192-007-1950-2

Bornmann, L., Weymuth, C., & Daniel, H.-D. (2010). A content analysis of referees' comments: How do comments on manuscripts rejected by a high-impact journal and later published in either a low- or high-impact journal differ? *Scientometrics, 83*, 493-506.

Boyer, E. L. (1997). Scholarship – A personal journey. In C. E. Glassick, M. T. Huber, & G. I. Maeroff (Eds.), *Scholarship assessed: Evaluation of the professoriate* (Special report). San Francisco, CA: Jossey-Bass.

Bozeman, B., & Boardman, C. (2009). Broad impacts and narrow perspectives: Passing the buck on science and social impacts. *Social Epistemology, 23*(3/4), 183-198. doi: 10.1080/02691720903364019

Braben, D. W. (2004). *Pioneering research: A risk worth taking*. Hoboken, NJ: Wiley.

Braben, D. W. (2011, October). *How to identify people who might radically change the way we think about an important subject*. Paper presented at the Danish National Research Foundation Annual Meeting, Copenhagen, Denmark. Retrieved from http://dg.dk/filer/20_aars_jubilaeum/Donald_Braben.pdf

The British Academy. (2007). *Peer review: The challenges for the humanities and social sciences*. Retrieved from http://www.britac.ac.uk/policy/peer-review.cfm

Brooks, J. H. (1988). Confidentiality of tenure review and discovery of peer review materials. *Brigham Young University Law Review, 1988*(4), 706-752.

Brown, R. S., & Kurland, J. E. (1990). Academic tenure and academic freedom. *Law and Contemporary Problems, 53*(3), 325-355. doi: 10.2307/1191800

Brown, T. (2004). *Peer review and the acceptance of new scientific idea*s. London, England: Sense about Science.

Burnham, J. C. (1990). The evolution of editorial peer review. *Journal of the American Medical Association, 263*(10), 1323-1329. doi: 10.1001/jama.263.10.1323

Burnham, J. C., Sauer, J. E., & Gibbs, R. D. (1987). Peer-reviewed grants in U.S. trade association research. *Science, Technology, & Human Values, 12*(2), 42-51.

doi:10.6342/NTU201602255

Cabezas-Clavijo, A., Robinson-García, N., Escabias, M., & Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? *PLoS ONE, 8*(6), e68258. doi: 10.1371/journal.pone.0068258

Caelleigh, A. S., Shea, J. A., & Penn, G. (2001). Selection and qualities of reviewers. *Academic Medicine, 76*(9), 914-916. doi: 10.1097/00001888-200109000-00016

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & De Mazancourt, C. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science, 338*(6110), 1065-1069.

Caligiuri, P., & Thomas, D.C. (2013). From the editors: How to write a high-quality review. *Journal of International Business Studies, 44*, 547-553.

Callaham, M. (2003). The evaluation and training of peer reviewers. In F. Godlee & T. Jefferson (Eds.), *Peer review in health science* (pp. 164-182). London, England: BMJ Publishing Group.

Callaham, M. (2005). The evaluation and training of peer reviewers. In F. Godlee & T. Jefferson (Eds.), *Peer review in health science* (pp. 164-182). London: BMJ Publishing Group.

Callaham, M. (2013). *What makes a good peer reviewer? The answer is not obvious*. Retrieved from https://www.elsevier.com/connect/what-makes-a-good-peer-reviewer-the-answer-is-not-obvious

Callaham, M. L, Baxt, W. G., Waeckerle, J. F., Wears, R. L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Journal of the American Medical Association, 280*(3), 229-230. doi: 10.1001/jama.280.3.229

Callaham, M. L., & Schriger, D. L. (2002). Effect of structured workshop training on subsequent performance of journal peer reviewers. *Annals of Emergency Medicine, 40*(3), 323-328. doi: 10.1067/mem.2002.127121

Callaham, M. L., Wears, R. L., & Waeckerle, J. F. (1998). Effect of attendance at a training session on peer reviewer quality and performance. *Annals of Emergency Medicine, 32*(3), 318-322. doi: 10.1016/S0196-0644(98)70007-1

Cameron, M. (2010). Faculty tenure in academe: The evolution, benefits and implications of an important tradition. *Journal of Student Affairs at New York University, 4*, 1-11.

doi:10.6342/NTU201602255

Campanario, J. M. (1998a). Peer review for journals as it stands today—Part 1. *Science Communication, 19*(3), 181-211. doi: 10.1177/1075547098019003002

Campanario, J. M. (1998b). Peer review for journals as it stands today—Part 2. *Science Communication, 19*(4), 277-306. doi: 10.1177/1075547098019004002

Chubin, D. E. (1994). Grants peer review in theory and practice. *Evaluation Review, 18*(1), 20-30. doi: 10.1177/0193841X9401800103

Chubin, D. E., & Hackett, E. J. (1990). *Peerless science: Peer review and U.S. science policy*. New York, NY: State University of New York Press.

Chubin, D. E., & Hackett, E. J. (2003, February). *Peer review for the 21st century: Applications to education research*. Paper presented at theWorkshop on the Peer Review of Education Research Grant Applications, Washington, DC.

Cicchetti, D. V. (1985). A critique of Whitehurst's "Interrater agreement for journal manuscript reviews": De omnibus, disputandem est. A*merican Psychologist, 40*(5), 563-568. doi: 10.1037/0003-066X.40.5.563

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*(1), 119-135. doi: 10.1017/S0140525X00065675

Cicchetti, D. V. (1997). Referees, editors, and publication practices: Improving the reliability and usefulness of the peer review system. *Science and Engineering Ethics, 3*(1), 51-62. doi: 10.1007/s11948-997-0016-4

Cole, J. R. (2000). The role of journals in the growth of scientific knowledge. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in honor of Eugene Garfield* (pp. 109-142). Medford, NJ: Information Today.

Cole, J. R., & Cole, S. (1973). *Social stratification in science*. Chicago, IL: University of Chicago Press.

Cole, S. (1992). *Making science: Between nature and society*. Cambridge, MA: Harvard University Press.

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science, 214*(4523), 881-886. doi: 10.1126/science.7302566

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505*(7485), 612-613. doi: 10.1038/505612a

doi:10.6342/NTU201602255

Cozzens, S. E. (1999). Are new accountability rules bad for science? *Issues in Science and Technology, 15*(4). Retrieved from http://issues.org/15-4/cozzens/

Cozzens, S. E. (2001). Autonomy and accountability for 21st century science. In J. de la Mothe (Ed.), *Science, technology, and governance* (pp. 104-115). London, England: Pinter.

Cummings, W. K., & Finkelstein, M. J. (2012). Historical and comparative perspectives on the faculty role in governance. In *Scholars in the changing American academy* (pp. 111-129). Dordrecht, Netherlands: Springer. doi: 10.1007/978-94-007-2730-4_8

Dalton, M. S. (1995). Refereeing of scholarly works for primary publishing. *Annual Review of Information Science and Technology, 30*, 213-250.

Daniel, H.-D. (1993). *Guardians of science. Fairness and reliability of peer review*. New York, NY: Wiley VCH.

Daniel, H.-D. (2005). Publications as a measure of scientific advancement and of scientists' productivity. *Learned Publishing, 18*, 143-148. doi: 10.1087/0953151053584939

De Vries, D. R., Marschall, E. A., & Stein, R. A. (2009). Exploring the peer review process: What is it, does it work, and can it be improved? *Fisheries, 34*(6), 270-279. doi: 10.1577/1548-8446-34.6.270

Dearing, R. (1997). *Higher education in the learning society* [Dearing report]. Leeds, England: National Committee of Inquiry into Higher Education.

Demicheli, V., & Di Pietrantonj, C. (2007). Peer review for improving the quality of grant applications. In *Cochrane database of systematic reviews* (Issue 2). Hoboken, NJ: Wiley. doi: 10.1002/14651858.MR000003.pub2

Dill, D., & Sporn, B. (1995). The implications of the postindustrial environment. In D. Dill & B. Sporn (Eds.), *Emerging patterns of social demand and university reform: Through a glass darkly* (pp. 1-19). Bingley, England: Emerald Group.

Donovan, C. (2011). State of the art in assessing research impact: Introduction to a special issue. *Research Evaluation, 20*(3), 175-179. doi: 10.3152/095820211X13118583635918

Drotar, D. (2009). Editorial: How to write effective reviews for the Journal of Pediatric Psychology. *Journal of Pediatric Psychology, 34*, 113-117.

doi:10.6342/NTU201602255

Eberley, S., & Warner, W. K. (1990). Fields or subfields of knowledge: Rejection rates and agreement in peer review. *American Sociologist, 21*(3), 217-231.

Eckberg, D. L. (1991). When nonreliability of reviews indicates solid science. *Behavioral and Brain Sciences, 14*(1), 145-146. doi: 10.1017/S0140525X00065791

Eckes, T. (2004). Rater agreement and rater severity: A many-faceted Rasch analysis of performance assessments in the "Test Deutsch als Fremdsprache" (TestDaF). *Diagnostica, 50*(2), 65-77.

Eisenhart, M. (2002). The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education, 32*(2), 241-255.

England, J. M. (1982). *A patron for pure science: The National Science Foundation's formative years, 1945-57*. Washington, DC: National Science Foundation.

Ernst, E., Saradeth, T., & Resch, K. L. (1993). Drawbacks of peer review. *Nature, 363*, 296. doi: 10.1038/363296a0

European Science Foundation. (2011a). *ESF survey analysis report on peer review practices*. Retrieved from http://www.esf.org/fileadmin/Public_documents/Publications/pr_guide_survey.pdf

European Science Foundation. (2011b). *European peer review guide: Integrating policies and practices into coherent procedures*. Retrieved from https://www.vr.se/download/18.2ab49299132224ae10680001647/European+Peer+Review+Guide.pdf

Evans, A. T., McNutt, R. A., Fletcher, S. W., & Fletcher, R. H. (1993). The characteristics of peer reviewers who produce good-quality reviews. *Journal of General Internal Medicine, 8*(8), 422-428. doi: 10.1007/BF02599618

Fagan, W. T. (1990). To accept or reject: Peer review. *Journal of Educational Thought, 24*(2), 103-113.

Fairweather, J. (2005). Beyond the rhetoric: Trends in the relative value of teaching and research in faculty salaries. *The Journal of Higher Education, 76*(4), 401-422. doi: 10.1353/jhe.2005.0027

Fang, F. C., & Casadevall, A. (2011). Retracted science and the retraction index. *Infection and Immunity, 79*(10), 3855-3859. doi: 10.1128/IAI.05661-11

Fang, H. (2011). Peer review and over-competitive research funding fostering

doi:10.6342/NTU201602255

mainstream opinion to monopoly. *Scientometrics, 87*, 293-301. doi: 10.1007/s11192-010-0323-4

Feurer, I. D., Becker, G. J., Picus, D., Ramirez, E., Darcy, M. D., & Hicks, M. E. (1994). Evaluating peer reviews: Pilot testing of a grading instrument. *Journal of the American Medical Association,* 272, 98-100.

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper – Diversity and uniqueness in reviewer comments. *American Psychologist, 45*, 591-598. doi: 10.1037/0003-066X.45.5.591

Fleiss, J. (1981). *Statistical methods for rates and proportions.* New York, NY: Wiley VCH.

Fletcher, R. H., & Fletcher, S. W. (2003). The effectiveness of journal peer review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health science* (pp. 62-75). London, England: BMJ Publishing Group.

Ford, E. (2013). Defining and characterizing open peer review: A review of the literature. *Journal of Scholarly Publishing, 44*(4), 311-326. doi: 10.3138/jsp.44-4-001

Friedman, D. P. (1995). Manuscript peer review at the *AJR*: Facts, figures, and quality assessment. *American Journal of Roentgenology, 164*(4), 1007-1009. doi: 10.2214/ajr.164.4.7726010

Frishauf, P. (2009). Reputation systems: A new vision for publishing and peer review. *Journal of Participatory Medicine, 1*(1). Retrieved from http://ojs.jopm.org/index.php/jpm/article/view/11

Frodeman, R., & Briggle, A. (2012). The dedisciplining of peer review. *Minerva, 50*(1), 3-19. doi: 10.1007/s11024-012-9192-8

Frodeman, R., & Holbrook, J. (2012, March). *The promise and perils of transformative research*. Paper presented at the Workshop on the Transformative research: Ethical and societal implications, National Science Foundation, Arlington, VA. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc84363

Frodeman, R., Holbrook, J., & Mitcham, C. (2012). Part I: Defining peer review. In R. Frodeman, J. Holbrook, C. Mitcham, & H. Xiaonan (Eds.), *Peer review, research integrity, and the governance of science: Practice, theory, and current discussions*. Beijing, China: People's Publishing House.

doi:10.6342/NTU201602255

Fuhrer, M. J., & Grabois, M. (1985). Grant application and review procedures of the National Institute of Handicapped Research: Survey of applicant and peer reviewer opinions. *Archives of Physical Medicine and Rehabilitation, 66*(5), 318-321.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*(3), 195-201. doi: 10.1002/asi.5090140304

Garfield, E., & Welljamsdorof, A. (1992). Citation data – Their use as quantitative indicators for science and technology evaluation and policy-making. *Current Contents, 49*, 5-13.

Garrow, J., Butterfield, M., Marshall, J., & Williamson, A. (1998). The reported training and experience of editors in chief of specialist clinical medical journals. *Journal of the American Medical Association, 280*(3), 286-287. doi: 10.1001/jama.280.3.286

Geisler, E. (2000). *The metrics of science and technology*. Westport, CT: Quorum Books.

Geisler, E. (2001). The mires of research evaluation. *The Scientist, 15*(10), 39.

General Accounting Office. (1999). *Federal research: Peer review practices at federal science agencies vary* (GAO/RCED-99-99). Washington, DC: United States General Accounting Office. Retrieved from http://science.energy.gov/~/media/bes/pdf/rc99099.pdf

Gibson, M., Spong, C. Y., Simonsen, S. E., Martin, S., & Scott, J. R. (2008). Author perception of peer review. *Obstetrics & Gynecology, 112*(3), 646-652. doi: 10.1097/AOG.0b013e31818425d4

Gillespie, G. W., Chubin, D. E., & Kurzon, G. M. (1985). Experience with NIH peer review: Researchers' cynicism and desire for change. *Science, Technology, & Human Values, 10*(3), 44-54. doi: 10.1177/016224398501000306

Gillett, R. (1993). Prescriptions for medical research II—Is medical research well served by peer review? *British Medical Journal, 306*(6893), 1672-1675. doi: 10.1136/bmj.306.6893.1672

Gilliland, S. W., & Cortina, J. M. (1997). Reviewer and editor decision making in the journal review process. *Personnel Psychology, 50*(2), 427-452.

Giraudeau, B., Leyrat, C., Le Gouge, A., Léger, J., & Caille, A. (2011). Peer review of

doi:10.6342/NTU201602255

grant applications: A simple method to identify proposals with discordant reviews. *PLoS ONE, 6*(11), e27557. doi: 10.1371/journal.pone.0027557

Global Research Council. (2012). *Statement of principles for scientific merit review*. Retrieved from http://www.globalresearchcouncil.org/sites/default/files/pdfs/gs_principles-English.pdf

Gluckman, P. (2012). *Which science to fund: Time to review peer review?* Auckland, New Zealand: Office of the Prime Minister's Science Advisory Committee. Retrieved from http://www.pmcsa.org.nz/wp-content/uploads/Which-science-to-fund-time-to-review-peer-review.pdf

Godlee, F., & Dickersin, K. (2003). Bias, subjectivity, chance, and conflict of interest. In F. Godlee & J. Jefferson (Eds.), *Peer review in health sciences* (2nd ed., pp. 91-117). London, England: BMJ Publishing Group.

Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. *Journal of the American Medical Association, 280*(3), 237-240. doi: 10.1001/jama.280.3.237

Goldman, R. L. (1994). The reliability of peer assessments: A meta-analysis. *Evaluation & the Health Professions, 17*, 3-21. doi: 10.1177/016327879401700101

Good, C. D., Parente, S. T., Rennie, D., & Fletcher, S. W. (1999). A worldwide assessment of medical journal editors' practices and needs - Results of a survey by the World Association of Medical Editors. *South African Medical Journal, 89*(4), 397-401.

Gordon, M. D. (1977). Evaluating the evaluators. *New Scientist, 73*, 342-343.

Gosden, H. (2003). 'Why not give us the full story?': Functions of referees' comments in peer reviews of scientific research papers. *Journal of English for Academic Purposes, 2*, 87-101. doi: 10.1016/S1475-1585(02)00037-1

Gottfredson, S. (1978). Evaluating psychological research reports. *American Psychologist, 33*(10), 920-934. doi: 10.1037//0003-066X.33.10.920

Greenbank, P. (2006). The academic's role: The need for re-evaluation? *Teaching in Higher Education, 11*(1), 107-112. doi: 10.1080/13562510500400248

doi:10.6342/NTU201602255

Gross-Schaefer, A., Gala, S., Jaccard, J., & Vetter, L. (2015). Being honest about tenure in the United States: The need for tenure system reform within institutions of higher education. *International Journal of Social Science Studies, 3*(4), 25-36. doi: 10.11114/ijsss.v3i4.827

Guston, D. H. (2003). The expanding role of peer review processes in the United States. In P. Shapira & S. Kuhlmann (Eds.), *Learning from science and technology policy evaluation: Experiences from the United States and Europe* (pp. 81-97). Cheltenham, England: Edward Elgar.

Guthrie, S., Guérin, B., Wu, H., Ismail, S., & Wooding, S. (2013). *Alternatives to peer review in research project funding* (RR-139-DH). Santa Monica, CA: Rand Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/research_reports/RR100/RR139/RAND_RR139.pdf

Hackett, E. J. (1997). Peer review in science and science policy. In M. S. Frankel (Ed.), *East-west dialogue on research evaluation in post-communist Europe* (pp. 51-60). Budapest, Hungary: Central European University Press.

Hames, I. (2007). *Peer review and manuscript management of scientific journals: Guidelines for good practice*. Oxford, England: Blackwell.

Harley, D., & Acord, S. K. (2011). *Peer review in academic promotion and publishing: Its meaning, locus, and future.* Berkeley, CA: Center for Studies in Higher Education, UC Berkeley.

Harley, D., Acord, S. K., Earl-Novell, S., Lawrence, S., & King, C. J. (2010). *Assessing the future landscape of scholarly communication: An exploration of faculty values and needs in seven disciplines.* Berkeley, CA: UC Berkeley, Center for Studies in Higher Education.

Harnad, S. (1985). Rational disagreement in peer review. *Science, Technology, & Human Values, 10*(3), 59. doi: 10.1177/016224398501000307

Harnad, S. (1996). Implementing peer review on the net: Scientific quality control in scholarly electronic journals. In R. Peek & G. Newby (Eds.), *Scholarly publishing: The electronic frontier* (pp. 103-118). Cambridge, MA: MIT Press.

Harnad, S. (2000). *The invisible hand of peer review*. Retrieved from http://cogprints.org/1646/

Harnad, S. (2008). *Peer review may not be perfect but alternatives are worse*. Retrieved

doi:10.6342/NTU201602255

from http://www.thetimes.co.uk/tto/news/uk/article1966879.ece

Hartmann, I., & Neidhardt, F. (1990). Peer review at the Deutsche Forschungsgemeinschaft. *Scientometrics, 19*(5-6), 419-425. doi: 10.1007/BF02020704

Healey, N. (2013, March 22). The problem with peer review. *Laboratory News*. Retrieved from http://www.labnews.co.uk/features/peer-review/

Heitman, E. (2002). The roots of honor and integrity in science: Historical themes in the practical ethics of research. In R. E. Bulger, E. Heitman, & S. J. Reiser (Eds.), *The ethical dimensions of the biological and health sciences* (pp. 21-28). Cambridge, England: Cambridge University Press.

Hemlin, S. (1996). Research on research evaluation. *Social Epistemology, 10*(2), 209-250. doi: 10.1080/02691729608578815

Hénard, F. (2010). *Learning our lesson: Review of quality teaching in higher education*. Paris, France: Organisation for Economic Cooperation and Development.

Hicks, D. M., & Katz, J. S. (1996). Where is science going? *Science, Technology, & Human Values, 21*(4), 379-406. doi: 10.1177/016224399602100401

The Higher Education Academy. (2009). *Reward and recognition in higher education: Institutional policies and their implementation*. Retrieved from https://www.heacademy.ac.uk/sites/default/files/rewardandrecognition_2_2.pdf

Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology, 50*(11), 1189-1195. doi: 10.1016/S0895-4356(97)00167-4

Hojat, M., Gonnella, J. B., & Caelleigh, A. (2003). Impartial judgment by the "gatekeepers" of science: Fallibility and accountability in the peer review process. *Advances in Health Sciences Education, 8*(1), 75-96. doi: 10.1023/A:1022670432373

Holbrook, J. B. (2010). The use of societal impacts considerations in grant proposal peer review: A comparison of five models. *Technology & Innovation, 12*(3), 213-224. doi: 10.3727/194982410X12895770314078

Holbrook, J. B. (2012). *Re-assessing the science – society relation: The case of the US National Science Foundation's broader impacts merit review criterion (1997 – 2011)*. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc77119/

doi:10.6342/NTU201602255

Holbrook, J. B. (2013a, October). *Peer review of team science research*. Paper presented at the Workshop on Institutional and Organizational Supports for Team Science, Washington, DC. Retrieved from http://hdl.handle.net/1853/49334

Holbrook, J. B. (2013b). What is interdisciplinary communication? Reflections on the very idea of disciplinary integration. *Synthese, 190*(11), 1865-1879. doi: 10.1007/s11229-012-0179-7

Holbrook, J. B., & Frodeman, R. (2011). Peer review and the *ex ante* assessment of societal impacts. *Research Evaluation, 20*(3), 239-246. doi: 10.3152/095820211X12941371876788

Holbrook, J. B., & Hrotic, S. (2013). Blue skies, impacts, and peer review. *Roars Transactions, a Journal on Research Policy and Evaluation, 1*(1). doi: 10.13130/2282-5398/2914

Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics, 79*(1), 171-190. doi: 10.1007/s11192-009-0411-5

Horrobin, D. F. (1990). The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association, 263*(10), 1438-1441. doi: 10.1001/jama.263.10.1438

Horrobin, D. F. (1996). Peer review of grant applications: A harbinger for mediocrity in clinical research. *The Lancet, 348*, 1293-1295. doi: 10.1016/S0140-6736(96)08029-4

Howard, L., & Wilkinson, G. (1998). Peer review and editorial decision-making. *The British Journal of Psychiatry, 173*(2), 110-113. doi: 10.1192/bjp.173.2.110

Huutoniemi, K., Klein, J. T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy, 39*(1), 79-88. doi: 10.1016/j.respol.2009.09.011

Ingelfinger, F. J. (1974). Peer review in biomedical publication. *American Journal of Medicine, 56*, 686-692. doi: 10.1016/0002-9343(74)90635-4

International Committee of Medical Journal Editors. (2013). *Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals*. Retrieved from http://www.scienceofsciencepolicy.net/sites/default/files/attachments/CAPR%20midterm.pdf

doi:10.6342/NTU201602255

Ismail, S., Farrands, A., & Wooding, S. (2009). *Evaluating grant peer review in the health sciences*. Cambridge, England: RAND Europe.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics, 69*(3), 591-606. doi: 10.1007/s11192-006-0171-4

Jefferson, T., Rudin, M., Brodney-Folse, S., & Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Methodology Reviews, 18*(2), No. MR000016. doi: 10.1002/14651858.MR000016.pub3

Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., Rennie, D., & PEER Investigators. (1998). Does masking author identity improve peer review quality? A randomized controlled trial. *Journal of the American Medical Association, 280*(3), 240-242. doi: 10.1001/jama.280.3.240

Kamenetzky, J. R. (2012). Opportunities for impact: Statistical analysis of the National Science Foundation's broader impacts criterion. *Science and Public Policy, 40*(1), 72-84. doi: 10.1093/scipol/scs059

Kassirer, J. P., & Campion, E. W. (1994). Peer review: Crude and understudied, but indispensable. *Journal of the American Medical Association, 272*(2), 96-97. doi: 10.1001/jama.272.2.96

Kostoff, R. N. (1995). Federal research impact assessment: Axioms, approaches, applications. *Scientometrics, 34*(2), 163-206. doi: 10.1007/BF02020420

Kostoff, R. N. (2004). *Research program peer review: Purposes, principles, practices, protocols*. Arlington, VA: Office of Naval Research.

Kreber, C. (2002). Controversy and consensus on the scholarship of teaching. *Studies in Higher Education, 27*(2), 151-167. doi: 10.1080/03075070220119995

Kronick, D. A. (1990). Peer review in 18th-century scientific journalism. *Journal of the American Medical Association, 263*(10), 1321-1322. doi: 10.1001/jama.263.10.1321

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist, 43*(8), 635-642. doi: 10.1037/0003-066X.43.8.635

Laband, D. N. (1990). Is there value-added from the review process in economics? Preliminary evidence from authors. *Quarterly Journal of Economics, 105*(2),

doi:10.6342/NTU201602255

341-352. doi: 10.2307/2937790

LaFollette, M. C. (1992). *Stealing into print: Fraud, plagiarism and misconduct in scientific publishing*. Berkeley, CA: University of California Press.

Lal, B., & Peña, V. (2013, March). *Big data in evaluating transformative scientific research: Concepts and a case study*. Paper presented at the Workshop on the Big Data: Measuring the Impact of the Government's Research and Development Investments, Washington, DC.

Laloë, F., & Mosseri, R. (2009). Bibliometric evaluation of individual researchers: Not even right… not even wrong! *Europhysics News, 40*(5), 26-29. doi: 10.1051/epn/2009704

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674054158

Langfeldt, L. (2001). The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science, 31*(6), 820-841. doi: 10.1177/030631201031006002

Langfeldt, L. (2006). The policy challenges of peer review: Managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation, 15*(1), 31-41. doi: 10.3152/147154406781776039

Langfeldt, L., & Kyvik, S. (2011). Researchers as evaluators: Tasks, tensions and politics. *Higher Education, 62*(2), 199-212. doi: 10.1007/s10734-010-9382-y

Laudel, G., & Glaser, J. (2012). *The ERC's impact on the grantees' research and their careers* (EURECIA, Work package 4 summary report). Retrieved from http://www.laudel.info/wp-content/uploads/2013/12/EURECIA-WP4-report-final-Jan2012.pdf

Ledin, A., Bornmann, L., Gannon, F., & Wallon, G. (2007). A persistent problem. *EMBO Reports, 8*(11), 982-987. doi: 10.1038/sj.embor.7401109

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2-17. doi: 10.1002/asi.22784

Levy, J. (1984). Peer review: The continual need for reassessment. *Cancer Investigation, 2*(4), 311-320. doi: 10.3109/07357908409018445

doi:10.6342/NTU201602255

Lock, S. (1985). *A difficult balance: Editorial peer review in medicine*. London, England: Nuffield Provincial Hospitals Trust.

Lovejoy, T. I., Revenson, T. A., & France, C. R. (2011). Reviewing manuscripts for peerreview journals: A primer for novice and seasoned reviewers. *Annals of Behavioral Medicine, 42*(1), 1-13.

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation, 21*, 48-60. doi: 10.1093/reseval/rvs001

Manske, P. T. (1997). A review of peer review. *Journal of Hand Surgery, 22A*(5), 767-771. doi: 10.1016/s0363-5023(97)80067-6

Marsh, H. W., & Bazeley, P. (1999). Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure. *Multivariate Behavioral Research, 34*(1), 1-30. doi: 10.1207/s15327906mbr3401_1

Marsh, H. W., Bond, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist, 42*, 33-38. doi: 10.1080/00050060600823275

Marsh, H., & Bornmann, L. (2009). Do women have less success in peer review? *Nature, 459*, 602. doi: 10.1038/nj7246-602a

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160-168. doi: 10.1037/0003-066X.63.3.160

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2011). Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis mode. *Journal of Informetrics, 5*, 167-181. doi: 10.1016/j.joi.2010.10.004

Martin, B. R. (2011). The research excellence framework and the "impact agenda": Are we creating a Frankenstein monster? *Research Evaluation, 20*(3), 247-254. doi: 10.3152/095820211X13118583635693

Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy, 12*(2), 61-90. doi: 10.1016/0048-7333(83)90005-7

McCullough, J. (1989). First comprehensive survey of NSF applicants focuses on their

doi:10.6342/NTU201602255

concerns about proposal review. *Science, Technology, & Human Values, 14*(1), 78-88. doi: 10.1177/016224398901400107

McNutt, R. A., Evans, A. T., Fletcher, R. H., & Fletcher, S. W. (1990). The effects of blinding on the quality of peer review: A randomized trial. *Journal of the American Medical Association, 263*(10), 1371-1376. doi: 10.1001/jama.1990.03440100079012

Merton, R. K. (1942). The normative structure of science. In N. W. Storer (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267-278). Chicago, IL: University of Chicago Press.

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis, 79*(4), 606-623. doi: 10.1086/354848

Mervis, J. (2011). Beyond the data. *Science, 334*(6053), 169-171. doi: 10.1126/science.334.6053.169

Metzger, W. P. (1990). The 1940 statement of principles on academic freedom and tenure. *Law and Contemporary Problems, 53*(3), 3. doi: 10.2307/1191793

Milem, J. F., Berger, J. B., & Dey, E. L. (2000). Faculty time allocation: A study of change over twenty years. *The Journal of Higher Education, 71*(4), 454-475. doi: 10.2307/2649148

Miller, D. A. (1978). Criteria for appointment, promotion, and retention of faculty in graduate social work programs. *Journal of Education for Social Work, 14*(2), 74-81. doi: 10.1080/00220612.1978.10671503

Modern Language Association. (2006). *Selected findings from the MLA's 2005 survey of tenure and promotion*. Retrieved from http://apps.mla.org/pdf/taskforcereportppt.pdf

Murphy, T. M., & Utts, J. M. (1994). A retrospective analysis of peer review at *Physiologia Plantarum*. *Physiologia Plantarum, 92*(3), 535-542. doi: 10.1034/j.1399-3054.1994.920401.x

Mutz, R., Bornmann, L., & Daniel, H.-D. (2015). Testing for fairness and predictive validity of research funding decisions: A multi-level multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the Austrian Science Fund. *Journal of the Association for Information Science and Technology, 66*(11), 2321-2339. doi: 10.1002/asi.23315

doi:10.6342/NTU201602255

National Institutes of Health. (2013). *Enhancing peer review survey results report.* Retrieved from http://enhancing-peer-review.nih.gov/docs/Enhancing_Peer_Review_Report_2012.pdf

National Science Foundation. (2011). *National Science Foundation's merit review criteria: Review and revisions.* Retrieved from http://www.nsf.gov/nsb/publications/2011/meritreviewcriteria.pdf

Nature. (2006). *Overview:* Nature*'s peer review trial*. Retrieved from http://www.nature.com/nature/peerreview/debate/nature05535.html

Neuman, W. L. (2000). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Allyn and Bacon.

Nickerson, R. S. (2005). What authors want from journal reviewers and editors. *American Psychologist, 60,* 661-662. doi: 10.1037/0003-066X.60.6.661

Nijstad, B. A. (2009). *Group performance*. East Sussex, NY: Psychology Press.

Nylenna, M., Riis, P., & Karlsson, Y. (1994). Multiple blinded reviews of the same two manuscripts: Effects of referee characteristics and publication language. *Journal of the American Medical Association, 272*(2), 149-151. doi: 10.1001/jama.272.2.149

Nylenna, M., Riis, P., & Karlson, Y. (1995). Are refereeing forms helpful? A study among medical referees in Denmark, Norway and Sweden. *European Science Editing, 55*, 3-5.

Odlyzko, A. M. (1996). Tragic loss or good riddance? The impending demise of traditional scholarly journals. In R. P. Peek & G. B. Newby (Eds.), *Scholarly publishing: The electronic frontier* (pp. 91-101). Cambridge, MA: MIT Press.

Office of Energy Efficiency and Renewable Energy. (2004). *Peer review guide*. Retrived from file:///E:/dissertation-20140904/2004peerreviewguide.pdf

Olbrecht, M., & Bornmann, L. (2010). Panel peer review of grant applications: What do we know from research in social psychology on judgment and decision-making in groups? *Research Evaluation, 19*(4), 293-304. doi: 10.3152/095820210X12809191250762

Opthof, T., & Wilde, A. A. M. (2009). The Hirsch-index: A simple, new tool for the assessment of scientific output of individual scientists: The case of Dutch professors in clinical cardiology. *Netherlands Heart Journal, 17*(4), 145-154. doi:

doi:10.6342/NTU201602255

10.1007/BF03086237

Organisation for Economic Co-operation and Development. (2011a). *Issue brief peer review*. Retrieved from http://www.oecd.org/innovation/policyplatform/48136766.pdf

Organisation for Economic Co-operation and Development. (2011b). *OECD issue brief: Research organization evaluation*. Retrieved from http://www.oecd.org/innovation/policyplatform/48136330.pdf

Over, R. (1996). Perceptions of the Australian research council large grants scheme: Differences between successful and unsuccessful applicants. *The Australian Educational Researcher, 23*(2), 17-36. doi: 10.1007/BF03219618

Oxman, A. D., Guyatt, G. H., Singer, J., Goldsmith, G. H., Hutchison, B. G., Milner, R. A., & Streiner, D. L. (1991). Agreement among reviewers of review articles. *Journal of Clinical Epidemiology, 44*, 91-98. doi: 10.1016/0895-4356(91)90205-N

Parliamentary Office of Science and Technology. (2002). Peer review. *Postnote, 182*, 1-4.

Pendlebury, D. A. (2008). *Using bibliometrics in evaluating research*. Philadelphia, PA: Thomson Scientific, Research Department.

Perper, T. (1989). The loss of innovation: Peer review in multi- and interdisciplinary research. *Issues in Integrative Studies, 7*, 21-56.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*(2), 187-195. doi: 10.1017/S0140525X00011183

Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citation scores – A case study in chemical engineering. *Journal of the American Society for Information Science, 45*, 39-49. doi: 10.1002/(SICI)1097-4571(199401)45:1<39::AID-ASI5>3.0.CO;2-Q

Polanyi, M. (1962). The republic of science. *Minerva, 1*(1), 54-73. doi: 10.1007/BF01101453

Popper, K. (1961). *The logic of scientific discovery*. London, England: Routledge & Kegan Paul.

Porter, R. (2005). What do grant reviewers really want, anyway? *Journal of Research Administration, 16*(2), 5-13.

doi:10.6342/NTU201602255

Pouris, A. (1988). Peer review in scientifically small countries. *R&D Management, 18*(4), 333-340. doi: 10.1111/j.1467-9310.1988.tb00608.x

Powell, K. (2010). Making the cut. *Nature, 467*, 383-385.

Pratt, D. (1997). Reconceptualising the evaluation of teaching in higher education. *Higher Education, 34*, 23-33.

Price, D. J. (1963). *Little science, big science*. New York, NY: Columbia University Press.

Rennie, D. (1986). Guarding the guardians: A conference on editorial peer review. *Journal of the American Medical Association, 256*(17), 2391-2392. doi: 10.1001/jama.256.17.2391

Rennie, D. (1998a). Freedom and responsibility in medical publication: Setting the balance right. *Journal of the American Medical Association, 280*(3), 300-302. doi: 10.1001/jama.280.3.300

Rennie, D. (1998b). The present state of medical journals. *The Lancet, 352*, S18-S22. doi: 10.1016/S0140-6736(98)90295-1

Rennie, D. (2003). Editorial peer review: Its development and rationale. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (pp. 1-13). London, England: BMJ Publishing Group.

Research Councils UK. (2006). *Report of the Research Councils UK efficiency and effectiveness of peer review project*. Retrieved from http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/rcukprreport.pdf

Research Councils UK. (2007). *RCUK response to the project report & consultation on the efficiency and effectiveness of peer review*. Retrieved from http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/responsereport.pdf

Research Evaluation and Policy Project. (2005). *Quantitative indicators for research assessment – A literature review* (REPP discussion paper 05/1). Canberra, Australia: Australian National University, Research School of Social Sciences, Research Evaluation and Policy Project.

Research Information Network. (2010). *Peer review: A guide for researchers*. Retrieved from http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/peer-re

doi:10.6342/NTU201602255

view-guide-researchers

Research Information Network. (2011). *E-journals: Their use, value and impact*. Retrieved from http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/e-journals-their-use-value-and-impact

Resnik, D. B., Gutierrez-Ford, C., & Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: An exploratory study. *Science and Engineering Ethics, 14*(3), 305-310. doi: 10.1007/s11948-008-9059-4

Rip, A. (2000). Higher forms of nonsense. *European Review, 8*(4), 467-485. doi: 10.1017/S1062798700005032

Roberts, M. R. (2009). Realizing societal benefit from academic research: Analysis of the National Science Foundation's broader impacts criterion. *Social Epistemology, 23*(3/4), 199-219. doi: 10.1080/02691720903364035

Robson, K., Pitt, L., & West, D. C. (2015). Navigating the peer review process: Reviewers' suggestions for a manuscript. *Journal of Advertising Research, 55*, 9-17. doi: 10.2501/JAR-55-1-009-017

Roy, R. (1985). Funding science: The real defects of peer-review and an alternative to it. *Science, Technology, & Human Values, 52*, 73-81. doi: 10.1177/016224398501000309

Royal Society. (1995). *Peer review - An assessment of recent developments*. Retrieved from https://royalsociety.org/~/media/Royal_Society_Content/policy/publications/1995/10260.pdf

Rubin, A., & Babbie, E. R. (2011). *Research methods for social work*. Belmont, CA: Thomson Brooks Cole.

Rymer, L. (2011). *Measuring the impact of research: The context for metric development*. Canberra, Australia: Group of Eight Australia.

Sandström, U., & Hällsten, M. (2007). Persistent nepotism in peer-review. *Scientometrics, 74*(2), 175-189. doi: 10.1007/s11192-008-0211-3

Sarewitz, D. (2004). How science makes environmental controversies worse. *Environmental Science & Policy, 7*(5), 385-403. doi: 10.1016/j.envsci.2004.06.001

Savage, J. D. (1999). *Funding science in America: Congress, universities, and the*

doi:10.6342/NTU201602255

politics of the academic pork barrel. New York, NY: Cambridge University Press.

Scarpa, T. (2009, May). *Assessing and advancing funding of biomedical research benchmarking: Values and practices of different countries*. Paper presented at the Sigtuna Project, Sigtuna, Sweden.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: Randomised controlled trial. *BMJ, 328*(7441), 673-678. doi: 10.1136/bmj.38023.700775.AE

Sense about Science. (2005). *"I don`t know what to believe ..." making sense of science stories*. Retrieved from http://www.senseaboutscience.org/data/files/resources/116/Embargoed_until_00.01Feb8th2013_IDKWTB_web.pdf

Sense about Science. (2010). *Peer review survey 2009: Full report*. Retrieved from http://www.senseaboutscience.org/data/files/Peer_Review/Peer_Review_Survey_Final_3.pdf

Shadbolt, N., Brody, T., Carr, L., & Harnad, S. (2006). The open research web: A preview of the optimal and the inevitable. In N. Jacobs (Ed.), *Open access: Key strategic, technical and economic aspects* (pp. 195-208). Oxford, England: Chandos. doi: 10.1016/B978-1-84334-203-8.50020-0

Shashok, K. (2005). Standardization vs diversity: How can we push peer review research forward? *Medscape General Medicine, 7*(1), 11.

Shashok, K. (2008). Content and communication: How can peer review provide helpful feedback about the writing? *BMC Medical Research Methods,8*(3).

Sieber, J. E. (2006). How can we research peer review? *Nature*. doi: 10.1038/nature05006

Siegelman, S. S. (1991). Assassins and Zealots: Variations in peer review (Special report). *Radiology, 178*(3), 637-642. doi: 10.1148/radiology.178.3.1994394

Sismondo, S. (1993). Some social constructions. *Social Studies of Science, 23*(3), 515-553. doi: 10.1177/0306312793023003004

Smith, L. C. (1981). Citation analysis. *Library Trends, 30*(1), 83-106.

Smith, R. (2003). The future of peer review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health science* (pp. 329-346). London, England: BMJ Publishing Group.

doi:10.6342/NTU201602255

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine.* 99, 178-182.

Smith, R. W. (2009). In search of an optimal peer review system. *Journal of Participatory Medicine, 1*, e13.

Smith, R., & Rennie, D. (1995). And now, evidence based editing. *BMJ, 311*(7009), 826-827. doi: 10.1136/bmj.311.7009.826

Snodgrass, R. (2006). Single- versus double-blind reviewing: An analysis of the literature. *Sigmod Record, 35*, 8-21. doi: 10.1145/1168092.1168094

Spencer, N. J., Hartnett, J., & Mahoney, J. (1986). Problems with reviews in the standard editorial practice. *Journal of Social Behavior and Personality, 1*(1), 21-36.

Spier, R. (2002a). The history of the peer-review process. *Trends in Biotechnology, 20*(8), 357-358. doi: 10.1016/S0167-7799(02)01985-6

Spier, R. (2002b). Peer review and innovation. *Science and Engineering Ethics, 8*, 99-108. doi: 10.1007/s11948-002-0035-0Steen, R. G. (2011). Retractions in the scientific literature: Do authors deliberately commit research fraud? *Journal of Medical Ethics, 37*, 113-117. doi: 10.1136/jme.2010.038125

Steen, R. G. (2011). Retractions in the scientific literature: do authors deliberately commit research fraud? Journal of medical ethics 37, 113–7.

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS ONE, 8*(7), e68397. doi: 10.1371/annotation/0d28db18-e117-4804-b1bc-e2da285103ac

Sternberg, R. J. (Ed.). (2005). *Reviewing scientific works in psychology.* Washington, DC: American Psychological Association.

Stossel, T. P. (1985). Reviewer status and review quality. *New England Journal of Medicine, 312*(10), 658-659. doi: 10.1056/NEJM198503073121024

Strayhorn, Jr., J., McDermont, J. F., & Tanguay, P. (1993). An intervention to improve the reliability of manuscript reviews for the *Journal of the American Academy of Child and Adolescent Psychiatry*. *American Journal of Psychiatry, 150*(6), 947-952. doi: 10.1176/ajp.150.6.947

Stricker, L. J. (1991). Disagreement among journal reviewers: No cause for undue alarm. *Behavioral and Brain Sciences, 14*(1), 163-164. doi:

doi:10.6342/NTU201602255

10.1017/S0140525X00065985

Tatum, C., & Wouters, P. (2013, November). *ACUMEN Portfolio: Resources for evaluation of individual researchers.* Paper presented at euroCRIS Membership Meeting, Porto, Portugal. Retrieved from http://tatum.cc/wp-content/uploads/Tatum-and-Wouters-ACUMEN@euroCRIS-Porto14Nov2013.pdf

Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values, 16*(3), 322-341. doi: 10.1177/016224399101600303

Trow, M. (1992). Thoughs on the White Paper of 1991. *Higher Education Quarterly, 46*(3), 213-216. doi: 10.1111/j.1468-2273.1992.tb01598.x

Tsang, E. W. K., & Frey, B. S. (2007). The as-is journal review process: Let authors own their ideas. *Academy of Management Learning and Education, 6*(1), 128-136. doi: 10.5465/AMLE.2007.24401710

Turcotte, C., Drolet, P., & Girard, M. (2004). Study design, originality and overall consistency influence acceptance or rejection of manuscripts submitted to the Journal. *Canadian Journal of Anaesthesia, 51*(6), 549-556. doi: 10.1007/BF03018396

Van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2014a). Academic talent selection in grant review panels. In K. Prpić, I. van der Weijden, & N. Asheulova (Eds.), *(Re)searching scientific careers* (pp. 25-54). St. Petersburg, Russia: IHST/RAS - Nestor-Historia – SSTNET/ESA.

Van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2014b). Different views on scholarly talent: What are the talents we are looking for in science? *Research Evaluation, 23,* 273-284. doi: 10.1093/reseval/rvu015

Van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review and project selection: A case study in the social and behavioral sciences. *Research Evaluation, 18*(4), 273-288. doi: 10.3152/095820209X475360

Van der Meulen, B., & Rip, A. (2000). Evaluation of societal quality of public sector research in the Netherlands. *Research Evaluation, 9*(1), 11-25. doi: 10.3152/147154400781777449

Van der Most, F. (2014). *ACUMEN Portfolio reveals a schism in academic evaluation*. Retrieved from https://reasonablyclose.net/wp/?p=197

doi:10.6342/NTU201602255

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397-420. doi: 10.1007/BF02129602

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*, 133-143. doi: 10.1007/s11192-005-0008-6

Van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology, 52*(7), 625-629. doi: 10.1016/S0895-4356(99)00047-5

Van Rooyen, S., Delamothe, T., & Evans, S. J. W. (2010). Effect on peer review of telling reviewers that their signed reviews might be posted on the web: Randomised controlled trial. *BMJ, 341*, c5729. doi: 10.1136/bmj.c5729

Van Rooyen, S., Godlee, F., Evans, S., Smith, R., & Black, N. (1998). Effect of blinding and unmasking on the quality of peer review: A randomized trial. *Journal of the American Medical Association, 280*(3), 234-237. doi: 10.1001/jama.280.3.234

Wager, L. (2008). Evidence-based editing: Promotion of editorial research. *Science Editor, 31*(1), 12.

Ware, M. (2013). *Peer review: An introduction and guide*. Retrieved from Publishing Research Consortium website: http://www.tandf.co.uk/journals/pdf/PRC-PeerReview-Guide-2013.pdf

Ware, M., & Monkman, M. (2008). *Peer review in scholarly journals: Perspective of the scholarly community – An international study*. Retrieved from http://publishingresearchconsortium.com/index.php/112-prc-projects/research-reports/peer-review-in-scholarly-journals-research-report/142-peer-review-in-scholarly-journals-perspective-of-the-scholarly-community-an-international-study

Waters, D. J. (2009, March 2). *Archives, edition-making, and the future of scholarly communication.* Retrieved from https://mellon.org/media/filer_public/30/9d/309de9a1-94fa-40fb-bb1f-f087333e8658/djw-archives-edition-making-2009.pdf

Weinbach, R. W., & Randolph, J. L. (1984). Ratings: Peer review for tenure and promotion in professional schools. *Improving College and University Teaching, 32*(2), 81-86. doi: 10.1080/00193089.1984.10533848

doi:10.6342/NTU201602255

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics, 62*(1), 117-131. doi: 10.1007/s11192-005-0007-7

Weiser, I. (2012). Peer review in the tenure and promotion process. *College Composition and Communication, 63*(4), 645-672.

Weller, A. C. (2002). *Editorial peer review: Its strengths and weaknesses*. Medford, NJ: American Society for Information Science and Technology.

Wenneras, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387*(6631), 341-343. doi: 10.1038/387341a0

Wessely, S. (1998). Peer review of grant applications: What do we know? *The Lancet, 352*(9124), 301-305. doi: 10.1016/S0140-6736(97)11129-1

Whitley, R., & Gläser, J. (Eds.) (2007). *The changing governance of the sciences: The advent of research evaluation systems* (Vols. 1-26). Dordrecht, Netherlands: Springer.

Wiener, S., Urivetsky, M., Bregman, D., Cohen, J., Eich, R., Gootman, N., … Wrigt, J. (1977). Peer review: Inter-reviewer agreement during evaluation of research grant evaluations. *Clinical Research, 25,* 306-311.

Wilcox, T. W. (1970). *A comprehensive survey of undergraduate programs in English in the United States*. Retrieved from http://files.eric.ed.gov/fulltext/ED044422.pdf

Wiley, S. (2008). Peer review isn't perfect ... But it's not a conspiracy designed to maintain the status quo. *The Scientist, 22*(11), 31.

Williams, J. (1977). Quality in the review process. *IEEE Transactions on Professional lCommunication, PC-20*(2), 131-132. doi: 10.1109/TPC.1977.6592349

Williamson, A. (2003). What will happen to peer review? *Learned Publishing, 16*(1), 15-20. doi: 10.1087/095315103320995041

Wong, V. S., & Callaham, M. L. (2012). Medical journal editors lacked familiarity with scientific publication issues despite training and regular exposure. *Journal of Clinical Epidemiology, 65*(3), 247-252. doi: 10.1016/j.jclinepi.2011.08.003

Wood, F. Q., & Wessely, S. (2003). Peer review of grant applications: A systematic review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (pp. 14-44). London, England: BMJ Publishing Group.

Wouters, P. (1997). Citation cycles and peer review cycles. *Scientometrics, 38*(1), 39-55.

doi:10.6342/NTU201602255

doi: 10.1007/BF02461122

Yalow, R. S. (1982). Competency testing for reviewers and editors. *Behavioral and Brain Sciences, 5*(2), 244-245. doi: 10.1017/S0140525X00011729

Young, S. N. (2003). Peer review of manuscripts: Theory and practice. *Journal of Psychiatry and Neuroscience, 28*(5), 327-330.

Young, P. (2006). Out of balance: Lecturers' perceptions of differential status and rewards in relation to teaching and research. *Teaching in Higher Education, 11*(2), 191-202. doi: 10.1080/13562510500527727

Ziman, J. (2000). *Real science: What it is and what it means*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511541391

Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva, 9*(1), 66-100. doi: 10.1007/BF01553188

doi:10.6342/NTU201602255

附錄一　期刊同儕審查評審標準分類架構（Bornmann, Nast, et al., 2008）

"Table S2: Zuordnung der criteria and reasons (n=542*) zu den neun Themen der Begutachtung (ggf. gemäß den Veränderungen von Frau Russon im Manuskript anpassen).
Innerhalb der Themen sind die criteria and reasons unter bestimmten key words zusammengefasst worden. Die Zahl hinter einem criterion or reason gibt die Identifikationsnummer der Studie an, aus der das criterion oder der reason entnommen wurde (siehe Tabelle S1 in den supporting information)."

(C) 'Relevance of contribution' (148 criteria and reasons)

Positive formulation (36 criteria and reasons)

Relevance of topic, in general (11 criteria and reasons)
Important, timely, relevant, critical, prevalent problem (5)
Topic is important (33)
Significant/ meaningful (39)
Interesting (39)
The topic selected was appropriate (21)
A controversial subject (31)
Social controversy surrounding topic (37, 38)
Valuable contribution (39)
Seminal piece of work/ research (31)
Needed/ in neglected field (39)
Breadth and appropriateness of topic (neither too broad nor too narrow) (2)

Relevance of topic to scientific advancement (9 criteria)
Substantive do'sd (evaluation of scientific improvement; e.g., it deals with an important topic; it attempts to unify the field) (20)
Where do we go from here (scientific advancement) (evaluation of general scientific relevance; e.g., it speaks to the central problems facing the discipline; it provokes much useful controversy; it outlines implications for future work) (20)
Researcher speaks to central problems facing the discipline (33)

doi:10.6342/NTU201602255

Importance or significance of topic (go beyond what is already known about the topic) (2)
Contribution: increment to the current literature (fills gaps in current knowledge) (6)
An advancement of knowledge (31)
Study represents a contribution to knowledge (33)
Originality/ heurism (evaluation of originality of research approach and of the heuristic value of research; e.g., it integrates data or findings from diverse sources into a coherent picture) (20)
Advances the discipline of radiology/ basic laboratory studies (36)

Originality, newness (8 criteria and reasons)
Originality/ heurism (evaluation of originality of research approach and of the heuristic value of research; e.g., it makes the reader think about something in a different way; it offers a new perspective on an old problem) (20)
New thought through extension (it is a "think-piece", an extension, elaboration or refinement of theory, and contains no new data) (32)
New/ novel treatment of subject (31)
A unique contribution (31)
Information original, in total or in part? (44)
Results provide new ideas for other researchers (33)
Research offers a new perspective on an existing problem (33)
Innovative (original observations, breaks new ground, seminal) (36)

Contribution to practical progress (5 criteria and reasons)
Practical, useful implications (5)
The paper provided useful information (21)
Study has practical implications (33)
Informative and useful (39)
Practical (impact on patient care, useful for the practicing radiologist, educational) (36)

Relevance of topic to journal (2 criteria and reasons)
Manuscript content: it is on the same topic as a number of articles recently published in the journal (24)
Topic of interest that differs in content (it is on a topic of interest to the field, but differs in content from articles traditionally published in the journal) (11, 24, 32)

Relevance of results (1 criterion)
Positive findings (3)

doi:10.6342/NTU201602255

Negative formulation (47 criteria and reasons)

Relevance of topic, in general (12 criteria and reasons)

Low priority (14)
Unimportant or irrelevant topic (5)
Content not important (30, 43)
Unimportant or insignificant (39)
Lack of importance and significance in the contribution (40)
Paper too specialised (22)
Content too technical (43)
Subject covered recently or scheduled for the future (30, 43)
Unimportant or insignificant contributions (4)
Magnitude of problem/ interest (e.g., the findings seem puzzling; it focuses exclusively on one side or aspect of the problem) (20)
General (e.g., contribution is of little or no importance) (16)
Probability of making at least moderate contribution if revised (19)

Relevance of topic to journal (12 criteria and reasons)

Not suitable for social problems (39)
Inappropriate for cjps (8), sss (9)
Inappropriate subject matter (10)
Inappropriate subject for journal (31)
Out of c&rl scope; little relevance to c&rl readership (21)
Out of scope or low value to the bulletin's audience (low substance, trivial, low priority, limited appeal, information not useful, too specialized, unduly theoretical and mathematical) (26)
The subject matter is not relevant for our audience (30)
Not relevant to journal (46)
Subject no interest to readers (31)
Unsuited to the readership (37)
Too technical for our pubilcation (30)
Direct replication/ replicating studies (it is a direct replication of an original study recently published in the journal; it adds no new dimension to theory)t (11, 32, 24)

Originality, newness (10 criteria and reasons)

Not new (14)
Lacks novelty/ redundant or obsolete findings (9)

doi:10.6342/NTU201602255

Newness of paper’s message (e.g., "old news", overworked) (26)
Contains nothing new (39)
Unoriginal (22)
Poor originality (46)
Idea not unique (30)
Old subject/ manuscript (31)
Offers little new material/ insights (21)
Manuscripts without new data: review (state-of-the-art) paper, contains no new data (24)

Relevance of topic to scientific advancement (7 criteria and reasons)
Manuscripts content: it is on a topic well outside the mainstream of the field (24)
Material well outside the mainstream (11)
Reviewer opinion: topic which most people in field consider important, but whose importance you believe to be greatly overemphasized (24)
Trivia (trivials) (notes about the insignificance of overall research; e.g., it contributes little or nothing to the field; the problem addressed is trivial) (20)
No significant addition to current body of knowledge (10)
Not relevant to the field (12)
Inadequate contribution to knowledge (8)

Relevance of results (4 criteria and reasons)
Resultsw (e.g., results are inconclusive, incomplete) (16)
Trivia (trivials) (notes about the insignificance of overall research; e.g., the results are trivial or unimportant) (20)
Statistically insignificant findings/ lack of statistical significance: study does not yield results which approach statistical significance (the theory tested is new and is the author's own)t (11, 24, 32)
Statistically insignificant findings/ lack of statistical significance (theory tested is one which is of current interest to the field)t (24, 32)

Contribution to practical progress (2 reasons)
Clinically not applicable (43)
Had too much or too little emphasis on practical implications (37)

doi:10.6342/NTU201602255

Neutral formulations (65 criteria and reasons)

Relevance of topic, in general (24 criteria and reasons)

Topic (1)
Topic selection (18, 42, 45)
Importance of topic: appropriateness (6)
Importance of conclusions (29)
Significance/ importance (44)
Content: wide and general interest? (13)
Content: extraordinary but special interest? (13)
Subject matter/ content (significance of the topic) (15)
Priority/ importance/ significance/ interest (17)
Appropriateness of topic (19)
Subject matter (22)
Significance of the paper (28, 29)
Significance of problem (29)
Relevance of subject (31)
Importance of subject (31)
Timeliness (31) of topic (37, 38)
Professional controversiality of topic (37, 38)
Overall priority for publication (34)
Contribution: overall contribution (theoretical, methodological, practical) (6)
Overall contribution (19)
Importance of the present contribution to the problem (35)
Discussion of educational issues (37)
Of interest to education professors (37)
Of interest to higher education professors (38)

Relevance of topic to scientific advancement (15 criteria and reasons)

Scientific importance of topic (41)
Relevance (5) to current areas of research (3)
Pertinence to current research in the discipline (7)
Contribution to knowledge (18, 42, 45)
Contribution to basic knowledge (37, 38)
Ho-hum research (the author uses precisely the same procedures as anyone else; the results are not overwhelming in their implications, but they constitute a needed component of knowledge; it is a continuation of previous studies by the same author) (20)

doi:10.6342/NTU201602255

Anticipatority of problems or issues in the field (37, 38)
The value of the author's findings to the advancement of the field (25, 27)
The potential contribution of the article to increased agricultural productivity (25)
Contribution to education as a field of study (37)
Contribution to higher education as a field of study (38)
Suggestions for future research/ thought/ action (37, 38)
Of long term definitive value (37, 38)
Contribution to field reviewed (41)
Suggestions for future research (42, 45)

Originality, newness (9 criteria and reasons)
Originality (3, 7, 34, 37, 38, 44)
Originality/ novelty (17)
The presence of original empirical evidence (25, 27)
Contribution: creativity and scope (6)
The creativity of ideas in the article (25, 27)
Previous presentation of the data at a conference (11)
Previous presentation at the regional/ national meetings/ at a conference, but not included in any proceedings (24, 32)
Previous presentation at the regional or national meetings with an abstract included in the proceedings (24)
Additional criteria for literature reviews and conceptual papers: uniqueness and incremental value (6)

Relevance of topic to journal (9 criteria and reasons)
Appropriateness (of manuscript, (41) for journal (17, 28, 29, 31)
The relevance of the article to the journal's focus (25, 27)
Relevance of manuscript to journal's readers (29)
Interest to readers (37, 38)
Probable interest of jpsp readers in the problem (assuming a sound paper) (35)
General interest (messages to a broad segment of the readership) (36)
Cja appropriate for this manuscript? (44)
Interest for a broad audience of psychologists (41)
Orientation to general rather than specialized readership (37, 38)

Contribution to practical progress (7 criteria and reasons)
Implications (15)
Importance of topic: practical importance (6)

doi:10.6342/NTU201602255

Practical implications (18, 45)
Contribution to practice (42)
Applicability to practical or applied problems (in the field) (3, 7, 37, 38)
The value of the article's findings to the affairs of everyday social life (27)
The value of the article's findings to clientel groups (25)

Relevance of results (1 reason)
Results (1, 5)

(W) 'Writing/ Presentation' (143 criteria and reasons)

Positive formulations (22 criteria and reasons)

Writing style and quality (13 criteria and reasons)
Well-written manuscript (clear, straightforward, easy to follow, logical) (5)
The paper was well written (21, 39)
Clarity/ coherence/ well written (31)
Good writing clarity and style (31)
Clarity of presentation and written expression (coherent and orderly presentation) (2)
Stylistic/ compositional do's (evaluation of description of research and results; e.g., it is well written; it avoids unrealistic speculation; the results are clearly presented) (20)
Well documented (solid documentation of conclusions, prospective studies, excellent gold standard, statstical analysis of results) (36)
Clear writing style makes it easy to assess the manuscript's quality (33)
Spirited style (37, 38)
Good taste (37, 38)
Professional appearance (31)
Professional style and tone (42)
Very thorough (31)

Quality of specific parts of manuscript (5 criteria and reasons)
High quality abstract (31)
Purpose of research is clearly stated (33)
Results are clearly presented (33)
Practical implications of study are made clear (33)
Problem well stated, formulated (5)

doi:10.6342/NTU201602255

Correctness (2 criteria)
Sentences are grammatically correct (33)
Sentences are properly punctuated (33)

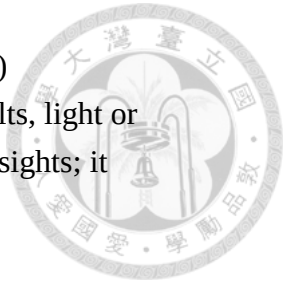
Organization/ length of manuscript (1 reason)
Deviation in length towards brevity (the manuscript is half as long as those smaller articles usually appearing in the journal (and for some reason cannot be treated as a "research note") (32)

Publication guidelines (1 reason)
Author guidelines followed (31)

Negative formulations (78 criteria and reasons)

Writing style and quality (32 criteria and reasons)
Poorly written (10, 21, 30, 37, 43)
Editorial and writing (e.g., generally poor writing) (16)
Poor writing or technical execution (37)
Written at wrong level (31)
Poor quality of writing (31)
Poorly written or presented (39)
Problems in presentation (poor writing, poor organization etc.) (4)
Quality of presentation (poorly written, padded, jargon, advertising or pr piece, too many quotes, doesn't follow bulletin style) (26)
Bad writing-clarity and style (31)
Writing style: incoherent, obscure, jargon, cluttered, bad tone (9)
Inappropriate writing style/ grammar (46)
Amateur style and tone (e.g., contrived emphasis - the frequent use of underlining or exclamation marks; exaggeration) (12)
Too esoteric (10)
Too speculativei (10)
Paper merely descriptive/ narrative (21)
Unprofessional appearance (31)
Too simple-'reporting'i (31)
The manuscript was too scholarly (37)
Unscholarly (37)
Poorly developed paper (21)

doi:10.6342/NTU201602255



Reads as a speech not an article, and would be difficult to change (30)
Data grinders; (evaluation of data processing; e.g., it is heavy on results, light or "spotty" on discussion of results; it contains more data, but no new insights; it emphasizes description rather than explanation)i (20)
Content inaccurate (43)
Readable (clumsy structure, awkward use of language) (36)
Content inaccurate or undocumented (30)
Content undocumented (43)
Text difficult to follow, to understand (5)
Unclear writing (authors don't appreciate the need of readers) (40)
Not thorough (31)
Inconsistent (39)
Undeveloped ideasl (e.g., premature, lack of focus, no stated purpose, no context, supperficial) (26)
Content not consistent with purpose (43)

Quality of specific parts of manuscript (28 criteria and reasons)
Insufficient definition - theory (authors did not provide definition, explanation, or reasoning for some of their variables, did not explain what the concepts mean) (12)
Purpose/ objective/ questions/ hypotheses unclear/ needed (21)
Designd (more information about subjects is needed) (16)
Methodology (lack of information) (15)
Statistics: inappropriate, incomplete, or insufficiently described, etc.m (5)
Statistical analysesd,m (e.g., unclear how an analysis was done) (16)
Proceduresm (e.g., procedural detail missing) (16)
Measurementm (e.g., reliability or validity data not given or unclear) (16)
Sampling method inappropriate or insufficiently describedm (5)
Inappropriate, suboptimal, insufficiently described instrumentm (5)
Intervention (independent variable) insufficiently described or confusingd (5)
Insufficient rationale - design (manuscripts lacked explanation of study procedures. No introduction to the true operational base of the research; e.g., describing the sample, saying who completed questionnaires, explain why the sample and procedure are appropriate to test the proposed research question) (12)
Subjects insufficiently described (5)
Conceptual: pre-executiond (e.g., concepts or issues not clearly described) (16)
Sample (ask for more information, size of the sample, (in terms of generalizability) (15)
Resultsc (e.g., tables and descriptive data are needed) (16)
Data presented with limited discussion of implicationsi (37)

doi:10.6342/NTU201602255

Incomplete, insufficient information in abstract (5)
Discuss/ elaborate a point (21)
Tables/ figures need clarification (21)
Unsatisfactory illustrations/ tables (46)
Macrostructure-organization and flow (e.g., the result section did not explicitly test each hypothesis raised in the theory section; the conclusion section might draw conclusions about theories and variables that were unrelated to the paper's explicit research question)i (12)
Conceptual: linkage to executiond (e.g., conceptualization or hypotheses unclear) (16)
Lack of relation between concepts and rationale of the study and methods or resultsd (40)
Case histories badly presented (30)
Nursing aspects not described well (30)
Poor statistics/ tables/ figures (31)
Title not representative of the study (5)

Organization/ length of manuscript (10 criteria and reasons)
Poor organization and presentation (revise/ resubmit) (awkward literary style, verbosity, redundancy etc.) (8)
Lacks organization; needs reorganization (21)
Organization/ style (e.g., connection between the theoretical background and the methodology is not clearly and logically explained; written in a rather informal style; sound intuitive rather than based in the literature) (15)
Too lengthy/ manuscript length (the manuscript is twice as long as those full-sized articles usually appearing in the journal and cannot be (intelligibly) condensed or divided) (11, 24)
Too long (39)
Paper too long/ short; delete/ add section (21)
Too short (39)
Manuscript length: the manuscript is half as long as those smaller articles usually appearing in the journal (and for some reason cannot be treated as a research note) (24)
Limited scope (revise to short communication) (8)
Written in an inappropriate format (22)

Correctness (5 criteria and reasons)
Errors in the article (39)
Technical (e.g., typographical errors) (15)
Defective tables or figures (5)

doi:10.6342/NTU201602255

Don't'sd (negative notations about quality; e.g., it misrepresents other viewpoints, literature, data etc.) (20)
Concepts poorly defined; terminology incorrectly used/ confusing (theoretical presentation incomplete, needs expansion): not well thought out (21)

Publication guidlines (3 reasons)
Editorial discretion: manuscript fails to meet criteria imposed at the discretion of the editor (4)
Failure to follow author guidelines (31)
Author guidelines not followed (31)

Neutral formulations (43 criteria and reasons)

Writing style and quality (17 criteria and reasons)
Style (1)
Writing (style), presentation (of manuscript) (5, 29)
Presentation: quality of writing (6)
Punctuation (18, 45)
Writing style and readability (18, 45)
Writing style (clarity) (19)
Clarity (17)
Clarity and conciseness of writing style (3, 7, 37, 38)
Quality of writing (organization, style, clarity) (28, 29)
Clarity, coherence, and conciseness of prose (41)
Additional criteria for literature reviews and conceptual papers: thoroughness (6)
Thoroughness (31)
Succinctness (35)
Presentation (44) level (42)
The entertainment quality of the essay (25, 27)
Emotional neutrality of the authori (37, 38)
Accuracy of information (41)

Quality of specific parts of manuscript (15 criteria and reasons)
Clarity of problem, hypothesis, and assumptions (29)
Clarity of purpose and problem definition (2)
Clarity of tables (45)/ of tabular material (18)
Abstract precise and complete? (44)
Sample and setting: sufficiency of description (6)

doi:10.6342/NTU201602255

Methods used are adequately described and appropriate?M (44)
Procedures: adequacy of description (6)
Data analysis and results: completeness (6)
Results clearly described? Tables, figures appropriate? (44)
Documentation of research design/ methodology (29)
Description of statistical analysis (29)
Title (of manuscript) (5, 31)
Title appropriate, informative? (44)
Abstract (5)
Problem statement (5)

Organization/ length of manscipt (9 criteria and reasons)
Organization of manuscript (41)
Style and organization of report (35)
Appropriateness of manuscript's total organization (37, 38)
Space (14)
Length (18) of manuscript (9) (appropriate to content?) (13)
Manuscript length (42, 45)
Form of the manuscript (text, figures, tables, nomenclature etc.) (13)
Should any portion be expanded, condensed, eliminated? (44)
Presentation/ appearance/ format (31)

Publication guidelines (2 criteria)
Presentation: conformance with publication guidelines (6)
Adherence to journal's stylistic guidelines (37, 38)

(D) 'Design/ Conception' (92 criteria and reasons)

Positive formulations (19 criteria and reasons)

Conceptual framework: Logic and correctness (8 criteria)
Research seems unbiased in research design (33)
Design - nonexperimental and cross-sectional: threat avoidance (uses needed control variables, logical time measurements etc.) (6)
Design - experimental and quasi-experimental: threat avoidance (minimizes and adresses threats to internal validity, statistical conclusion validity, construct validity etc.) (6)

doi:10.6342/NTU201602255

Study contains no internal contradictions or computational errors (33)
Study integrates findings from diverse sources into a coherent picture (33)
Based on current research in field or research tradition (37, 38)
Balance and fairness in coverage of alternative views (41)
Design - meta-analysis: incremental value (goes beyond simply summarizing the data, explores moderators fully) (6)

Quality and appropriateness (6 criteria and reasons)
Good design (39)
Design of study is adequate (33)
Well-designed study (appropriate, rigorous, comprehensive design) (5)
Research was well executed (33)
Thorough and complete (39)
Breadth (37) (wide coverage) (38)

Quality of sampling (2 reasons)
Sample size sufficiently large (5)
Good sampling (39)

Generalizability (2 criteria)
Substantive do'sc (evaluation of scientific improvement; e.g., it has excellent generalizability) (20)
Study has wide generalizability (33)

Replicability (1 criterion)
Replicability of the review (being able to arrive at the same conclusions as the author) (2)

Negative formulations (40 criteria and reasons)

Conceptual framework: Logic and correctness (20 criteria and reasons)
Conceptual: pre-executionw (e.g., conceptual basis for study poor or incomplete) (16)
Insufficient or incomplete problem statement (5)
Intervention (independent variable) insufficiently described or confusingw (5)
Lack of conceptual or theoretical frameworkt (5)
Lack of relation between concepts and rationale of the study and methods or resultsw (40)
Concepts and operationalizations not in alignment (the operational base of research did

doi:10.6342/NTU201602255

not reflect the variables or models under study) (12)
Substantive ommissions/ naivete (9)
Designw (design is defective, incomplete or inappropriate) (16)
Scientific validity: faulty design (too few cases, poor response rate, inappropriate methodology for problem adressed, factually inaccurate, shifting base for statistical calculations) (26)
Theory or concepts incorrectly or inadequately usedt (39)
Potential confounding variables not addressedm (5)
Statistical analysesm,w (e.g., rationale or justification for the analysis is faulty) (16)
Experimental data with no control group (32) (the study is a field experiment but contains no control goup) (11)
Design and analysis characteristics: the study is an experiment but contains no control group (24)
Control problem (experiment) (21)
Key issues not addressed (21)
Missing literature or variablesl (23)
Conceptual: linkage to executionw (e.g., experiment or task does not test the theory) (16)
Descriptive orientation (37)
Too superficial (10)

Quality and appropriateness (10 criteria and reasons)
Research poorly organized (39)
Poorly designed (39)
Poor research design (43)
Slight, trivial, or low quality work/ research (31)
Low quality (14)
Based upon poor research design or faulty methodologym (30)
Inadequate research (10) design (study lacked validity) (12)
Inappropriate experimental design/ statisticsm (8)
Don't'sw (negative notations about quality; e.g., the problem has not been considered carefully enough; the design used does not justify the conclusions drawn) (20)
Case reports (22)

Quality of sampling (6 criteria and reasons)
Sample too small or biased (5)
Sampling problem (21)
Sampling inadequate or incorrect (39)

doi:10.6342/NTU201602255

Design and analysis characteristics: study is based in a sample size of less than 30 (24)
Insufficient data (one year, on site, one cultivar, one replication)m (8)
Magnitude of problem/ interestc (e.g., it presents a small amount of data from a large research project) (20)

Generalizability (4 criteria and reasons)
Study generalizability and interpretability: it is a laboratory study and gives no evidence of generalizability to other samples or situations (24)
Study generalizability and interpretability: it is a one-company field study and gives no evidence of generalizability to other companies (24)
Pilot study research, with little evidence of generalizability (11)
Narrow scope; lacks generalizability (21)

Neutral formulations (33 criteria and reasons)

Quality and appropriateness (16 criteria and reasons)
Design (1) of study (18)
Research design (5, 42, 45)
Quality of research (design and analysis)m (28, 29)
Adequacy of research design (19) and analysism (35)
Appropriateness of research design/ methodsm (29)
Experimental design (34)
Experimental design appropriate? (44)
Use of experimental as opposed to non-experimental designs (37)
Design - experimental and quasi-experimental: appropiateness (6)
Design - nonexperimental and cross-sectional: appropriateness (6)
Design - meta-analysis: procedural adequacy (6)
Design - qualitative: procedural adequacy (6)
Completeness of coverage (41)
Scientific (44)
Descriptive-orientated (case studies) (38)
Soundness/ quality (17)

Conceptual framework: Logic and correctness (9 criteria)
Design - experimental and quasi-experimental: proper controls (comparison groups) (6)
Design - experimental and quasi-experimental: valid manipulations (6)
The grasp of the author's research design on the question investigated (25, 27)
Logical rigori (3, 7, 42)

doi:10.6342/NTU201602255

Conceptualisation (31)
Conceptual development: adequacy of scope and complexity (6)
Conceptual development: clarity and logical coherence (6)
Conceptual or theoretical argumentst (19)
Do data verify hypotheses and conclusions?I (13)

Quality of sampling (5 criteria and reasons)
Sample (1, 15) and sampling (5)
Sample and setting: appropriateness (6)
Sample and setting: justifications (6)
Design - meta-analysis: adequacy of sample of studies (6)
Appropriateness of population, sampling, and data gathering techniques (29)

Replicability (2 criteria)
Replicability (3) (if research article) (37, 38)
Replicability of research (42) techniquesm (7)

Generalizability (1 criterion)
Generalizability and validity of results (31)

(M) 'Method/ Statistics' (72 criteria and reasons)

Positive formulations (8 criteria and reasons)

Correctness and appropriateness (3 criteria and reasons)
Analyzing interval data properly (it contains interval data and is treated accordingly) (11)
Accurate statistical data (32)
Methods are adequate to test the research questions (33)

Method/ statistics: In general (2 reasons)
Good methodology (39)
Good analysis (39)

doi:10.6342/NTU201602255

Newness (2 reasons)

Novel, unique approach to data analysis (5)

New statistical methods/ including data techniques (it discusses a new statistical test or a new data-collection technique, and contains no new data) (11, 32)

Quality of operationalization and measurement (1 criterion)

Operational definitions are adequate (33)

Negative formulations (29 criteria and reasons)

Correctness and appropriateness (19 criteria and reasons)

Inappropriate experimental design/ statisticsd (8)

Inappropriate procedures/ methodology (8)

Application of inappropriate analyses (32) (parametric tests for ordinal data; the sample is fairly large) (11)

Statistics: inappropriate, incomplete, or insufficiently described, etc.w (5)

Poor statistical analysis (46)

Methodological shortcomings or flaws (criticize data for being inappropriate or insufficient to test the hypothesis presented) (4)

Methodology (study: limitations because of methodology) (15)

Sampling method inappropriate or insufficiently describedw (5)

Overengineering (overdoing of methodology; exotic and sophisticated statistical techniques to analyze data) (12)

Study generalizability and interpretability: correlation analysis only, which does not permit cause-effect inferences to be drawn (24)

Inappropriate, suboptimal, insufficiently described instrumentw (5)

Statistics inadequately handled (39)

Potential confounding variables not addressedd (5)

Defective methodology: sampling, generalizability, measurement errors (9)

Statistical analysesw,d (e.g., rationale or justification for the analysis is faulty) (16)

Proceduresw (e.g., confounding, nonindependence, or lack of counterbalancing) (16)

Design and analysis characteristics: author uses parametric statistical tests, although his data are ordinal (sample is fairly large) (24)

Misunderstanding/ misapplication of data and literature, referee incredulityl (9)

Based upon poor research design or faulty methodologyd (30)

doi:10.6342/NTU201602255

Quality of operationalization and measurement (6 reasons)

Measurementw (e.g., measure is indirect, superficial, not the best) (16)
Inaccurate or inconsistent data reported (5, 46)
Insufficient data (46) presented (5)
Insufficient data (one year, on site, one cultivar, one replication)d (8)
Data poorly interpreted or insufficienti (39)
Scores insufficiently reliable or unknown reliability (5)

Method/ statistics: In general (4 reasons)

Methodological problems (14)
Poor methodologically/ methodology (22, 46)
Methodology poor or incorrect (39)
Poor analysis (39)

Neutral formulations (35 criteria and reasons)

Correctness and appropriateness (14 criteria and reasons)

Scholarly or systematic approach to review (systematic and organized approach to the review of literature; methodology has to "fit" with the research question) (2)
Data analysis and results: appropriateness of statistics (6)
The sophistication of the author's research methodology and data analysis (25, 27)
Adequacy of research design and analysisd (35)
Quality of research (design and analysis)d (28, 29)
Appropriateness of research design/ methodsd (29)
Appropriateness of statistical analysis (29)
Data analysis (appropriateness) (44)
Appropriate use of statistics (37) / appropriate use of statistic (if research article) (38)
Methods of data analysis appropriate? (44)
Methods used are adequately described and appropriate?W (44)
Use of standard empirical methodologies (37)
Data analysis and results: warranted assumptions and appropriate error rates (6)
Replicability of research (42) techniquesd (7)

Methodology/ statistics: In general (10 criteria and reasons)

Data (1)
Data type (nominal, ordinal, interval; treated accordingly) (24)
Analysis (1)
Data analysis and interpretationi (19)

doi:10.6342/NTU201602255

Statistical analyses (18, 42, 45)
Procedures: quality (6)
Methodology (15)
Mathematical precision (7)
Math/ stat rigor (3)
Research method (31)

Quality of operationalization and measurement (10 criteria and reasons)
Measurement (1)
Measurements: operationalization (6)
Operationalization of key constructs (19)
Instrumentation and data collection (5)
Measurements: procedural adequacy (6)
Measurements: availability (6)
Measurements: reliability (6)
Measurements: validity (6)
Methodology (reliability of the data) (15)
Adequacy of reliability and validity of measures (29)

Newness (1 criterion)
Manuscripts without new data: it discusses a new statistical test or a new data collection technique and contains no new data (24)

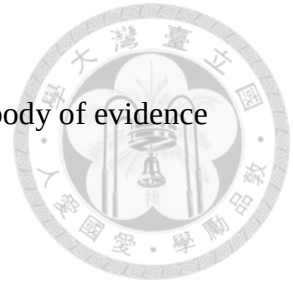
(I) 'Discussion of results' (46 criteria and reasons)

Positive formulations (10 criteria and reasons)

Correctness, adequacy, and objectivity (5 criteria)
Objectivity in reporting results (18, 45)
Data support the conclusions (34)
Conclusions properly based in the results (29)
Results justify the conclusions drawn (33)
Results are objectively reported and interpreted (33)

Clarity (3 criteria)
Existence of and persuasiveness in arguing for a well-articulated point of view (41)

doi:10.6342/NTU201602255



Existence and clarity of a take-home message (41)
Clarity of conclusions (conclusions are precise and follow from the body of evidence reviewed) (2)

Breath of interpretation (2 criteria and reasons)
Interpretation took into account the limitations of the study (5)
Depth (intensive examination of specific area) (38)

Negative formulations (17 criteria and reasons)

Correctness, adequacy, and objectivity (14 reasons)
Underinterpretation of results, ignoring results (5)
Overinterpretation of the results (5, 46)
Overgeneralized (39)
Inadequate interpretation (8)
Poor argumentation: superficial analysis, opinion only, polemical, atheoretical, unscholarly (9)
Too speculativew (10)
Too simple-'reporting'w (31)
Data poorly interpreted or insufficientm (39)
Flaws in the logic (23)
Scientific validity: faulty conclusions (26)
Conclusions not in alignement (12)
Interpretations and conclusions (e.g., statement is unacceptable, unconvincing) (16)
Macrostructure-organization and flow (e.g; the conclusion section might draw conclusions about theories and variables that were unrelated to the paper's explicit research question)w (12)
Interpretations/ conclusions not warranted by data (21) as compiled (30)

Breath of interpretation (3 criteria)
Data presented with limited discussion of implicationsw (37)
Data grinders; (evaluation of data processing; e.g., it is heavy on results, light or "spotty" on discussion of results; it contains more data, but no new insights; it emphasizes description rather than explanation)w (20)
Absence of a message (46)

doi:10.6342/NTU201602255

Neutral formulations (19 criteria and reasons)

Correctness, adequacy, and objectivity (8 criteria and reasons)
Logical rigord (3, 7, 42)
Validity of logic used (37, 38)
Reasonableness of conclusions (17)
Do data verify hypotheses and conclusions?D (13)
Conclusions appropriate in relation to experimental design and results? (44)
Discussion appropriate and complete? (44)
Design - qualitative: appropriateness of conclusions (6)
Emotional neutrality of the authorw (37, 38)

Discussion of results, in general (5 criteria and reasons)
Discussion and conclusion (5)
Data analysis and interpretationm (19)
Discussion of data implications (37)
Discussion of limitations of data or theory (presented) (37, 38)
Discussion and conclusion: explanation of results (6)

Breath of interpretation (4 criteria)
Discussion and conclusion: derivation of implications (6)
Discussion and conclusion: description of limitations (6)
Development of alternative interpretation of data presented (37, 38)
Depth (37)

Clarity (2 criteria)
Clarity of conclusions/ generalizations (29)
Clarity of arguments (2)

(L) 'Basis of Literature' (27 criteria and reasons)

Positive formulations (3 criteria and reasons)

Coverage of relevant literature (3 criteria and reasons)
Thoughtful, focused, up-to-date review of the literature (5)
Literature review is thourough and up-to-date (33)

doi:10.6342/NTU201602255

References: many references to earlier publications in the same journal (24)

Negative formulations (10 criteria and reasons)

Coverage of relevant literature (7 criteria and reasons)

Ignorance of relevant literature (9)
Technical (e.g., omission on references in the bibliography, bibliographical style) (15)
Body of literature omitted (21)
Missing literature or variablesd (23)
Ignorant of previously published work on the same subject (39)
References: no reference to earlier publications in the same journal (24)
Undeveloped ideasw (e.g., previous work not investigated) (26)

Literatur review: Correctness (3 reasons)

Inadequate, incomplete, inaccurate, or outdated review of the literature (5)
Misunderstanding/ misapplication of data and literature, referee incredulitym (9)
Amateur style and tone (e.g., negative approach to the previous literature) (12)

Neutral formulations (14 criteria and reasons)

Coverage of relevant literature (10 criteria and reasons)

Literature review: thoroughness and accuracy (6)
Review of literature to date on the subject (38)
Literature review: linkage to most important literature (6)
Coverage of literature (3)/ adequacy of coverage of literature 2)
Coverage of significant existing literature (7, 42)
Mastery of relevant literature (19)
Attention to relevant literature (35)
Documentation of relationship to previous research (29)
All relevant references cited? Excessive number cited? (44)
Existence or nonexistence of references to earlier publications in the same journal (11, 32)

Basis of literature, in general (4 criteria and reasons)

Review of literature (1, 18, 37)
Literature review (45)
Literature review: framing within the literature (6)
Use of bibliography (37, 38)

doi:10.6342/NTU201602255

(T) 'Theory' (24 criteria and reasons)

Positive formulations (5 criteria and reasons)

Newness, interest of theory (5 criteria and reasons)

Manuscript representing a new, original theory (the theory tested is new and the author's own) (11)

New theory and the authors own (24)

New, original theory with statistically significant results (the theory tested is new and is the author's own) (32)

Study proposes a new theory to explain existing research findings (33)

Research is of theoretical interest and importance (33)

Negative formulations (11 criteria and reasons)

Contribution/ importance to theory (6 criteria and reasons)

Direct replication/ replicating studies (it is a direct replication of an original study recently published in the journal; it adds no new dimension to theory)c (11, 32, 24)

Reviewer opinion: theory which most people in the field are interested in, but which you consider to be method-bound (24)

Manuscripts without new data: a "think-piece", an extension, elaboration, or refinement of theory, and contains no new data (24)

Reviewer opinion: theory which most people in the field are interested in, but which you consider to be flawed or erroneous (24)

Statistically insignificant findings/ lack of statistical significance: study does not yield results which approach statistical significance (the theory tested is new and is the author's own)c (11,24, 32)

Statistically insignificant findings/ lack of statistical significance: study does not yield results which approach statistical significance; the theory tested is one which is of current interest to the field)c (24, 32)

Theory, in general (5 reasons)

Lack of conceptual or theoretical frameworkd (5)

No theory (little or no theory to explain relationships amoung variables) (12)

Holes in the theory (23)

Theory or concepts incorrectly or inadequately usedd (39)

Theoretical problems: theoretical framework is unsound (4)

doi:10.6342/NTU201602255

Neutral formulations (8 criteria and reasons)

Theory, in general (5 criteria and reasons)
Theory (1, 42)
Theoretical model (18, 45)
Theoretical orientation (37, 38) of manuscript (41)
Theoretically grounded (38)
Conceptual or theoretical argumentsd (19)

Contribution/ importance to theory (3 criterion)
Importance of topic: theoretical importance (6)
Theoretical significance (3, 7)
The theoretical relevance of the question (25) investigated (27)

(A) 'Authors Reputation/ Institutional Affiliation' (11 criteria and reasons)

Positive formulations (3 criteria and reasons)

Reputation, affiliation (3 criteria and reasons)
Author's reputation (you know who the author is and believe that s/he has a justifiably strong reputation in the area s/he writes about) (11, 24, 32)
Author being a member of the journals advisory board (11, 24, 32)
Author's affiliation: author is a member of the company/ university which sponsors the journal (24)

Negative formulations (2 criteria and reasons)

Reputation (2 criteria and reasons)
Author appears to heave weak or inappropriate credentials for the subject matter (10)
Author reputation (you know who the author is and believe that he has no reputation in the area he writes about) (11, 24)

doi:10.6342/NTU201602255

Neutral formulations (6 criteria)

Reputation, affiliation (6 criteria)
Reputation (42) of author (18)
Institutional affiliation (18, 45)
The background and reputation of the author (25, 27)
Reputation of the author or institution (37, 38)
Author's status and reputation (45)
The scholarship demonstrated in the article (25, 27)

(E) 'Ethics' (10 criteria and reasons)

Positive formulations (0 criteria or reasons)

No criterion or reason

Negative formulations (5 criteria and reasons)

Multiple publication (4 criteria and reasons)
Cutting up the data (paper under review for one journal overlapped by 80% a paper under review for another journal) (12)
Previously published elsewhere (30)
Prior publication (the paper was submitted elsewhere) (26)
Previous presentation at the regional/ national meetings and fully included in the proceedings (24, 32)

Secondary analysis (1 criterion and reason)
Only secondary analysis of data presented by others (it contains only a secondary analysis of data previously collected and analyzed by others) (11, 24)

Neutral formulations (5 criteria and reasons)

Disciplinary ethics (5 criteria and reasons)
Scientific ethics (42)
Any concerns about ethics? (44)
Compatibility with generally accepted disciplinary ethics (7)
The ethical sense demonstrated by the author (25, 27)

doi:10.6342/NTU201602255

Compatability with disciplinary ethics (37, 38)

Not assignable (13 criteria and reasons)

Disciplinary hubris/ program differences (9)
Other (10, 21)
Ad hominem (1)
General (1)
Reviews: appropriate review articles (36)
Rsna meeting papers (presented at the annual meeting of rsna) (36)
Policy oriented (position paper)/ value oriented (opinion pieces) (37, 38)
Reviewers' comments (14)
Recommendation (13, 35, 44) regarding publishing (34)
Acceptance/ rejection (17)
Overall evaluation of manuscript (in its present form) (29)
Likely overall evaluation once author(s) has made modifications (those that are possible and likely do be done) (29)
Contribution: publication potential (likely to improve contribution substantially with revision of article; has strengths in some parts that offsets weaknesses) (6)

* die Summe aller criteria and reasons, die den neun Themen der Begutachtung zugeordnet wurden, ergibt 586 (42 criteria and reasons wurden zwei Themen, ein reason drei Themen zugeordnet). Falls ein criterion or reason mehrfach zugeordnet wurde, ist dies am criterion or reason vermerkt.

doi:10.6342/NTU201602255

附錄二　評審報告總字數統計量分析

一、統計量綜合比較

		總統計量	量化研究	非量化研究	社會科學研究	人文科學研究
個數	有效的	103	43	60	43	60
	遺漏值	0	0	0	0	0
平均數		703.02	549.16	813.28	633.84	752.60
中位數		556.00	392.00	663.00	416.00	619.00
眾數		157[a]	201	157	201[a]	157[a]
標準差		653.210	552.368	700.601	615.020	680.004
範圍		4900	2862	4900	2862	4900
最小值		40	66	40	66	40
最大值		4940	2928	4940	2928	4940
總和		72411	23614	48797	633.84	752.60
百分位數	25	320.00	201.00	410.50	392.75	392.75
	50	556.00	392.00	663.00	619.00	619.00
	75	896.00	699.00	1055.25	931.75	931.75

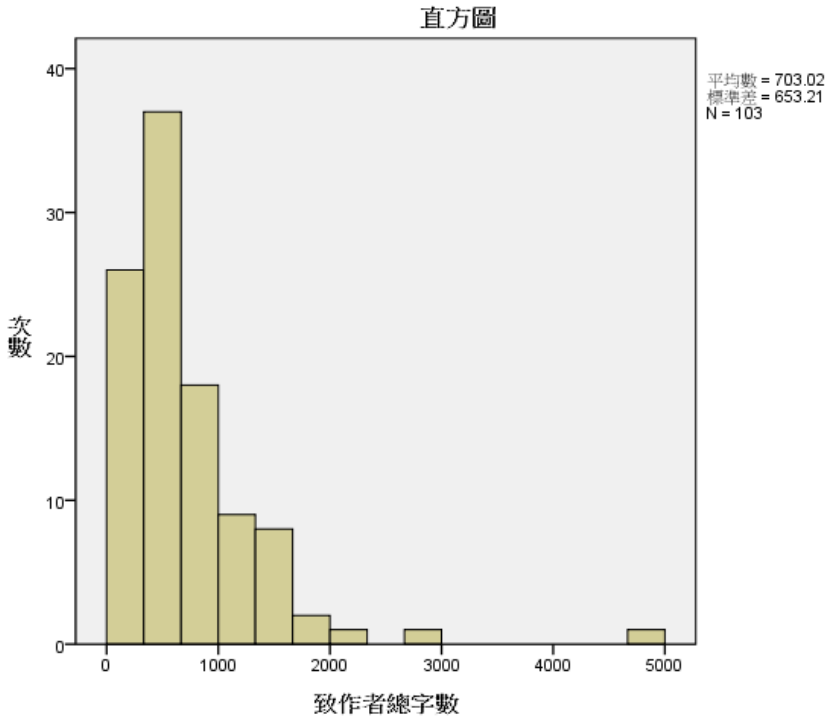
a. 存在多個眾數，顯示的為最小值。

doi:10.6342/NTU201602255

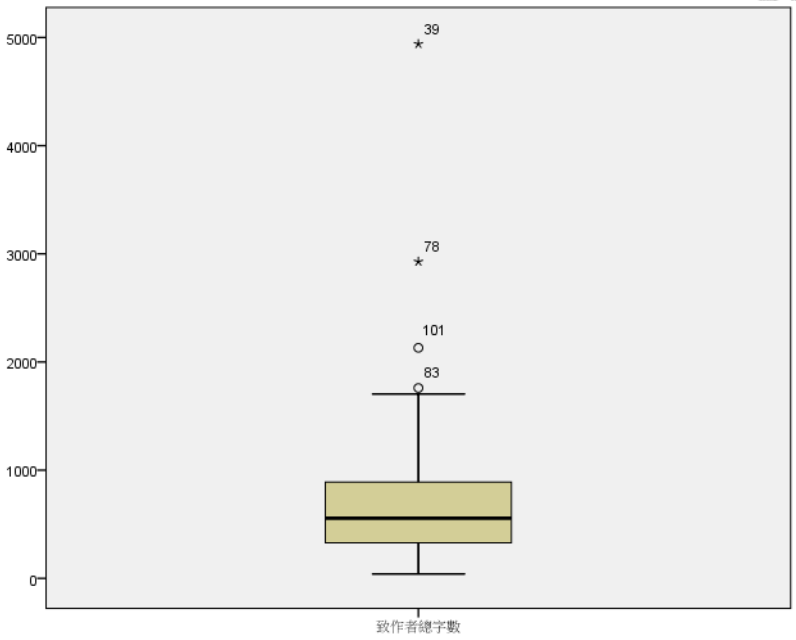
二、總統計量

描述性統計量

			統計量	標準誤
致作者總字數	平均數		703.02	64.363
	平均數的 95% 信賴區間	下限	575.36	
		上限	830.68	
	刪除兩極端各 5% 觀察值之平均數		629.14	
	中位數		556.00	
	變異數		426683.509	
	標準差		653.210	
	最小值		40	
	最大值		4940	
	範圍		4900	
	四分位全距		576	
	偏態		3.290	.238
	峰度		17.391	.472



doi:10.6342/NTU201602255



致作者總字數 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
6.00	0 .	469999
10.00	1 .	0334555688
7.00	2 .	0001268
13.00	3 .	0123333455799
13.00	4 .	0011111466679
7.00	5 .	1357888
9.00	6 .	123556699
6.00	7 .	025778
7.00	8 .	1445789
3.00	9 .	113
3.00	10 .	026
1.00	11 .	8
4.00	12 .	3344
3.00	13 .	288
5.00	14 .	12337
1.00	15 .	0
.00	16 .	
1.00	17 .	0
4.00 Extremes		(>=1760)

Stem width: 100

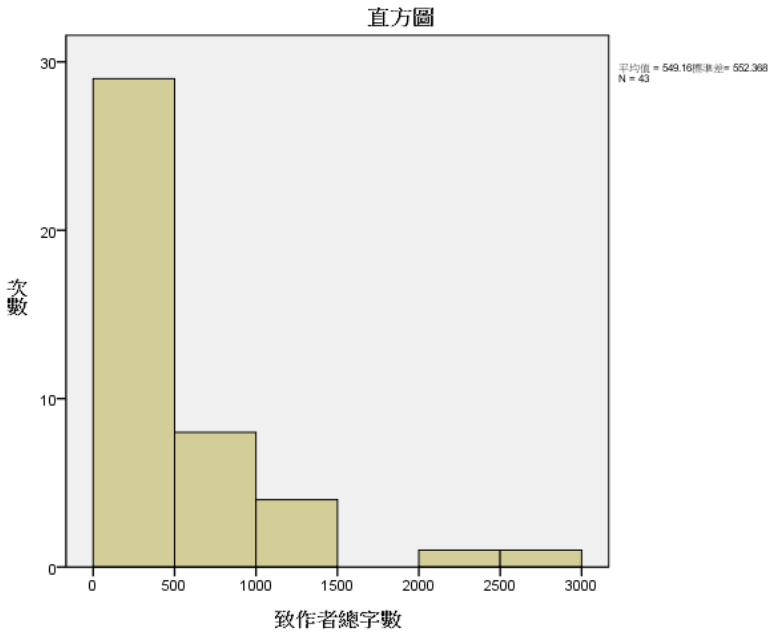
Each leaf: 1 case(s)

doi:10.6342/NTU201602255

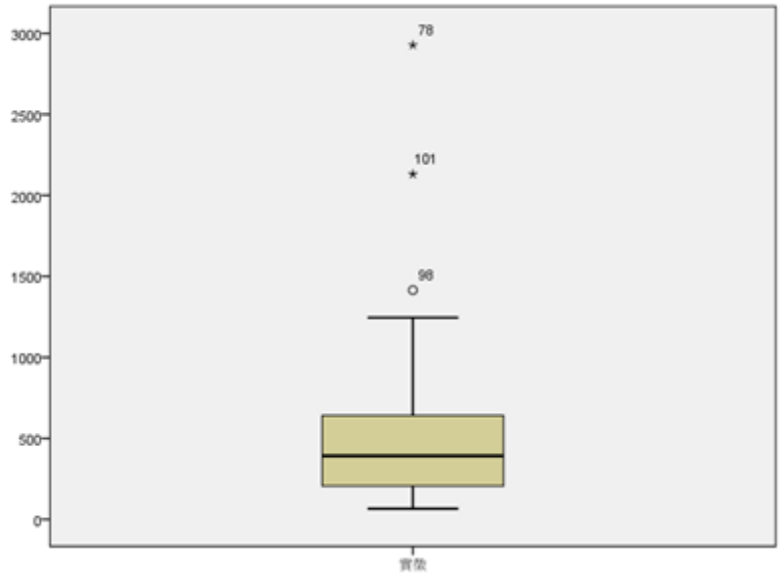
三、量化研究

描述性統計量

		統計量	標準誤
致作者總字數	平均數	549.16	84.235
	平均數的 95% 信賴區間 下限	379.17	
	上限	719.16	
	刪除兩極端各 5% 觀察值之平均數	469.56	
	中位數	392.00	
	變異數	305110.854	
	標準差	552.368	
	最小值	66	
	最大值	2928	
	範圍	2862	
	四分位全距	498	
	偏態	2.663	.361
	峰度	8.495	.709



doi:10.6342/NTU201602255



致作者總字數 Stem-and-Leaf Plot for
量化研究

Frequency	Stem &	Leaf
3.00	0 .	699
5.00	1 .	03468
5.00	2 .	00016
9.00	3 .	013334579
7.00	4 .	0114666
3.00	5 .	138
1.00	6 .	9
3.00	7 .	257
1.00	8 .	4
.00	9 .	
1.00	10 .	0
.00	11 .	
2.00	12 .	44
3.00 Extremes		(>=1415)

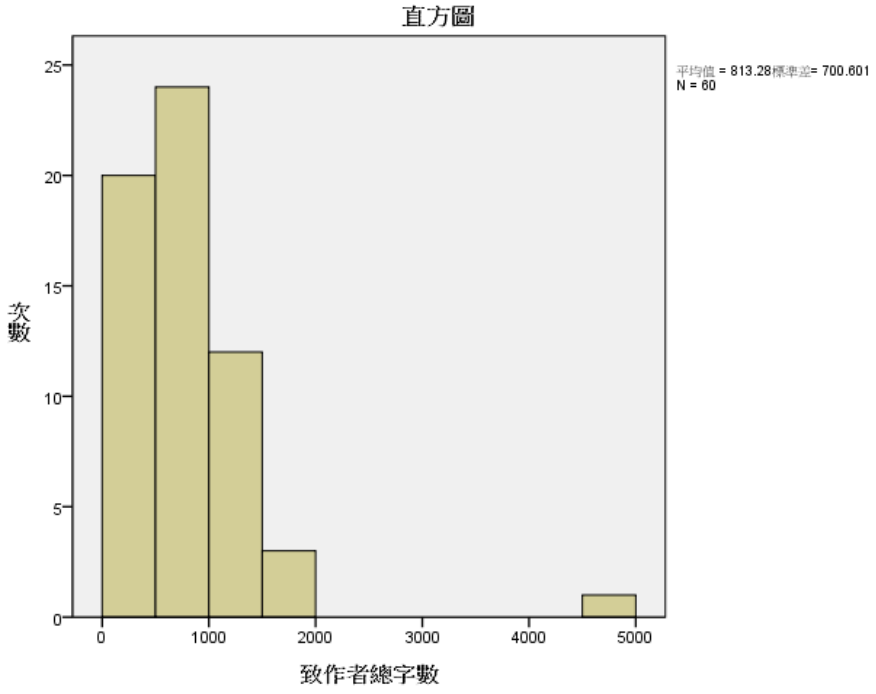
Stem width: 100
Each leaf: 1 case(s)

doi:10.6342/NTU201602255

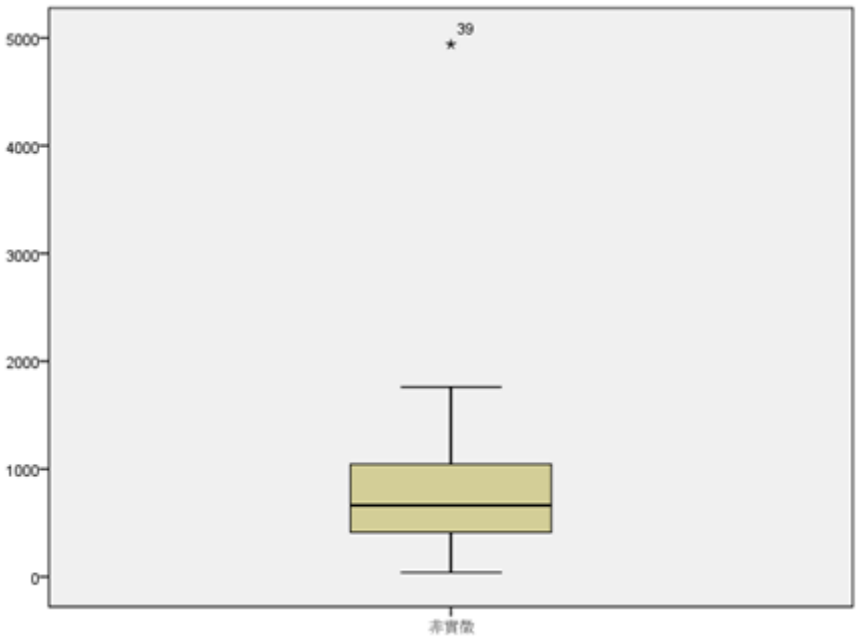
四、非量化研究

描述性統計量

		統計量	標準誤
致作者總字數	平均數	813.28	90.447
	平均數的 95% 信賴區間 下限	632.30	
	平均數的 95% 信賴區間 上限	994.27	
	刪除兩極端各 5% 觀察值之平均數	743.76	
	中位數	663.00	
	變異數	490842.105	
	標準差	700.601	
	最小值	40	
	最大值	4940	
	範圍	4900	
	四分位全距	645	
	偏態	3.576	.309
	峰度	19.820	.608



doi:10.6342/NTU201602255



致作者總字數 Stem-and-Leaf Plot for
非量化研究

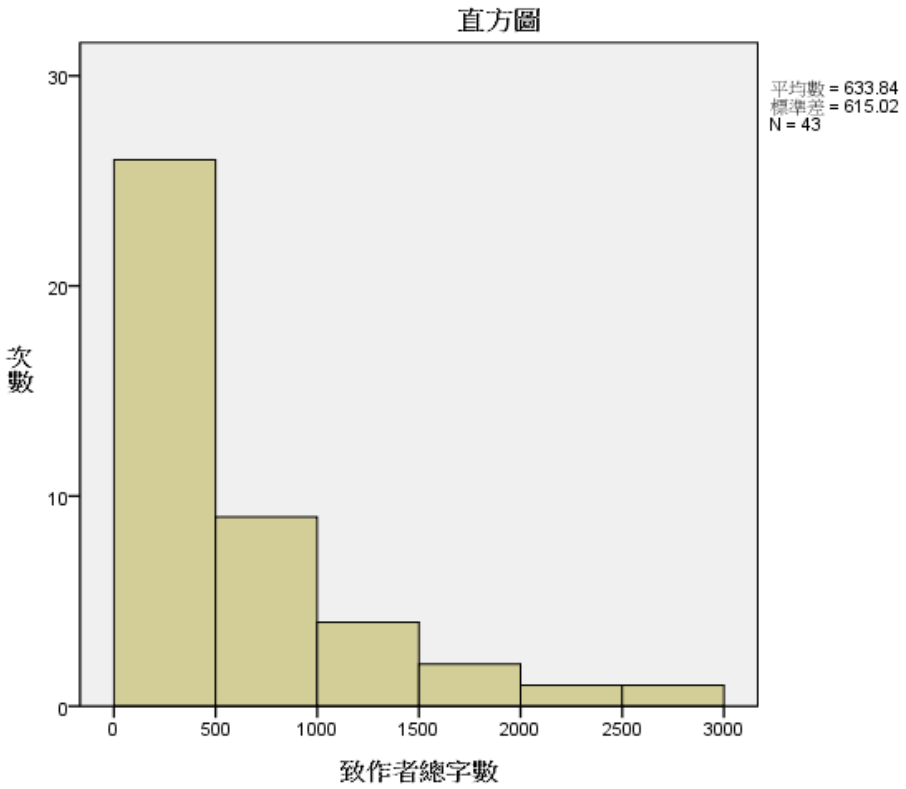
Frequency	Stem &	Leaf
8.00	0 .	00011111
6.00	0 .	223333
10.00	0 .	4444445555
11.00	0 .	66666666777
9.00	0 .	888888999
3.00	1 .	001
5.00	1 .	22333
5.00	1 .	44445
2.00	1 .	77
1.00 Extremes		(>=4940)

Stem width: 1000
Each leaf: 1 case(s)

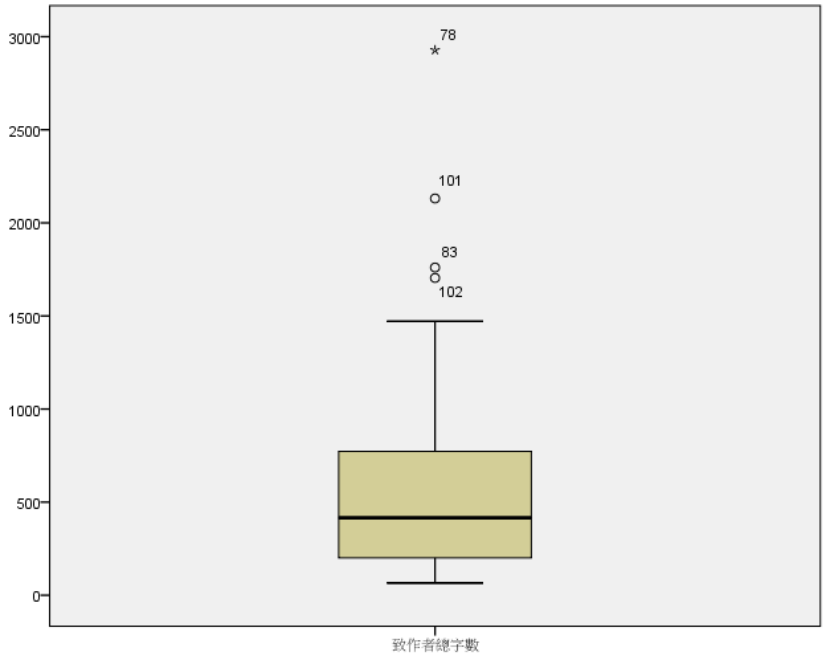
doi:10.6342/NTU201602255

描述性統計量

			統計量	標準誤
致作者總字數	平均數		633.84	93.790
	平均數的 95% 信賴區間	下限	444.56	
		上限	823.11	
	刪除兩極端各 5% 觀察值之平均數		562.31	
	中位數		416.00	
	變異數		378249.711	
	標準差		615.020	
	最小值		66	
	最大值		2928	
	範圍		2862	
	四分位全距		573	
	偏態		1.900	.361
	峰度		3.973	.709



doi:10.6342/NTU201602255



致作者總字數 Stem-and-Leaf Plot

社會科學研究

Frequency	Stem &	Leaf
9.00	0 .	000111111
11.00	0 .	22222333333
7.00	0 .	4444445
6.00	0 .	667777
2.00	0 .	89
1.00	1 .	0
1.00	1 .	2
2.00	1 .	44
4.00 Extremes		(>=1704)

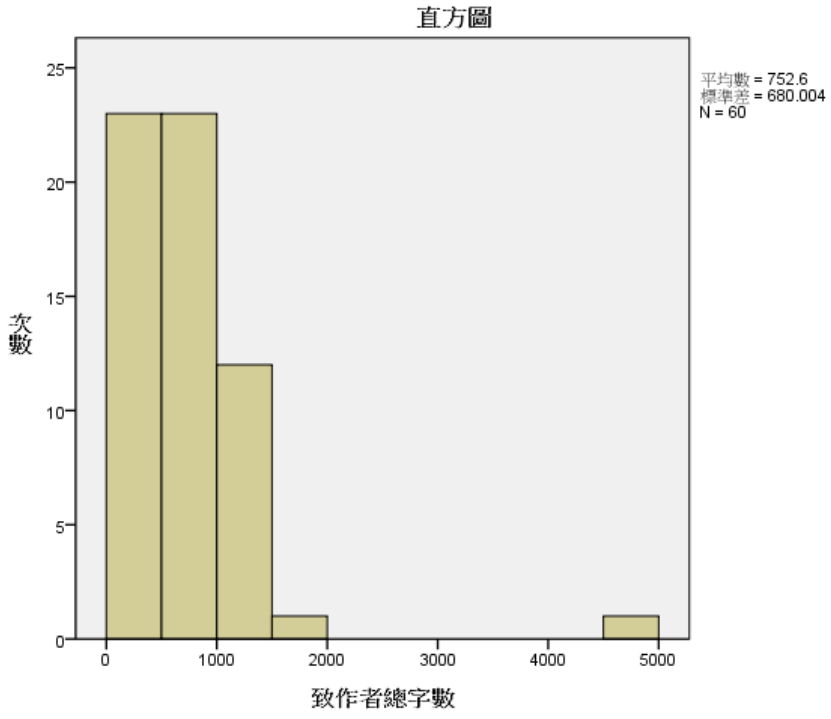
Stem width: 1000

Each leaf: 1 case(s)

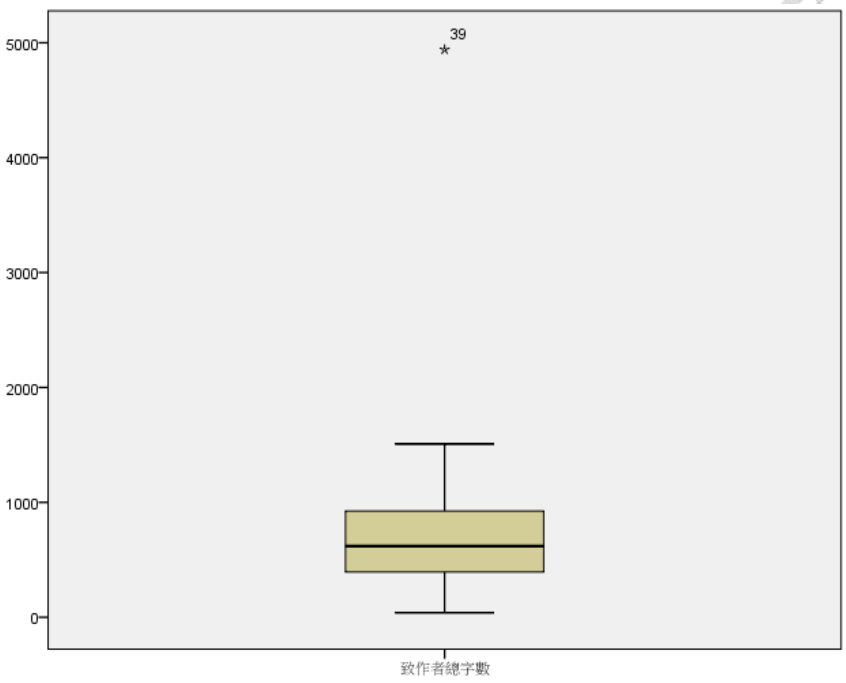
doi:10.6342/NTU201602255

描述性統計量

			統計量	標準誤
致作者總字數	平均數		752.60	87.788
	平均數的 95% 信賴區間	下限	576.94	
		上限	928.26	
	刪除兩極端各 5% 觀察值之平均數		685.89	
	中位數		619.00	
	變異數		462405.600	
	標準差		680.004	
	最小值		40	
	最大值		4940	
	範圍		4900	
	四分位全距		539	
	偏態		4.084	.309
	峰度		24.100	.608



doi:10.6342/NTU201602255



致作者總字數 Stem-and-Leaf Plot

人文科學研究

Frequency	Stem &	Leaf
7.00	0 .	0001111
9.00	0 .	223333333
13.00	0 .	4444444555555
9.00	0 .	666666677
8.00	0 .	88888899
3.00	1 .	001
6.00	1 .	222333
4.00	1 .	4445
1.00 Extremes		(>=4940)

Stem width: 1000

Each leaf: 1 case(s)

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255

附錄三　評審者之評語筆數統計量分析

一、統計量綜合比較

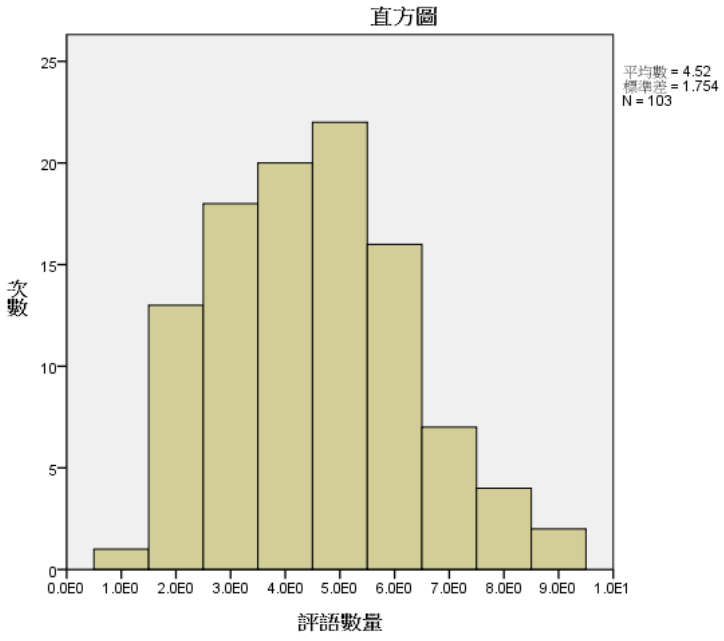
		總統計量	量化研究	非量化研究	社會科學研究	人文科學研究
個數	有效的	103	43	60	43	60
	遺漏值	0	0	0	0	0
平均數		4.52	4.88	4.27	5.14	4.08
中位數		4.00	5.00	4.00	5.00	4.00
眾數		5	6	5	6	5
標準差		1.754	1.991	1.528	1.971	1.441
範圍		8	8	6	7	7
最小值		1	1	2	2	1
最大值		9	9	8	9	8
百分位數	25	3.00	3.00	3.00	3.00	3.00
	50	4.00	5.00	4.00	5.00	4.00
	75	6.00	6.00	5.00	6.00	5.00

doi:10.6342/NTU201602255

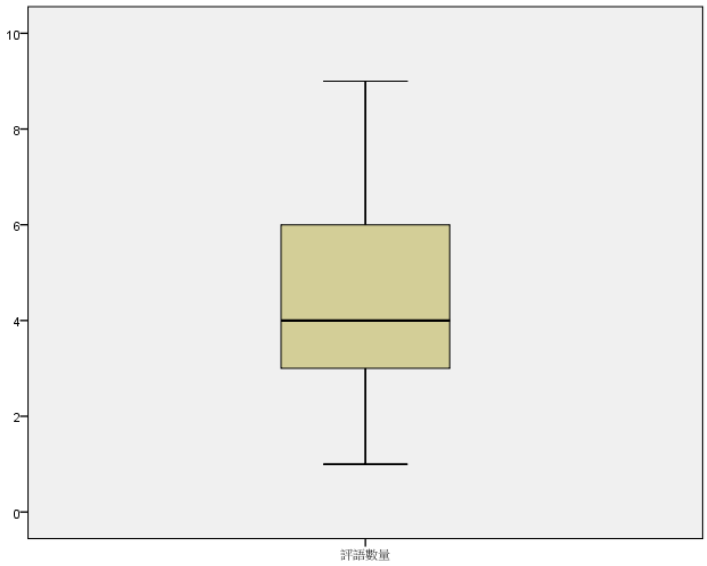
二、總統計量

描述性統計量

			統計量	標準誤
評語數量	平均數		4.52	.173
	平均數的 95% 信賴區間	下限	4.18	
		上限	4.87	
	刪除兩極端各 5% 觀察值之平均數		4.46	
	中位數		4.00	
	變異數		3.075	
	標準差		1.754	
	最小值		1	
	最大值		9	
	範圍		8	
	四分位全距		3	
	偏態		.348	.238
	峰度		-.334	.472



doi:10.6342/NTU201602255



評語數量 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	1 .	0
.00	1 .	
13.00	2 .	0000000000000
.00	2 .	
18.00	3 .	000000000000000000
.00	3 .	
20.00	4 .	00000000000000000000
.00	4 .	
22.00	5 .	0000000000000000000000
.00	5 .	
16.00	6 .	0000000000000000
.00	6 .	
7.00	7 .	0000000
.00	7 .	
4.00	8 .	0000
.00	8 .	
2.00	9 .	00

Stem width: 1

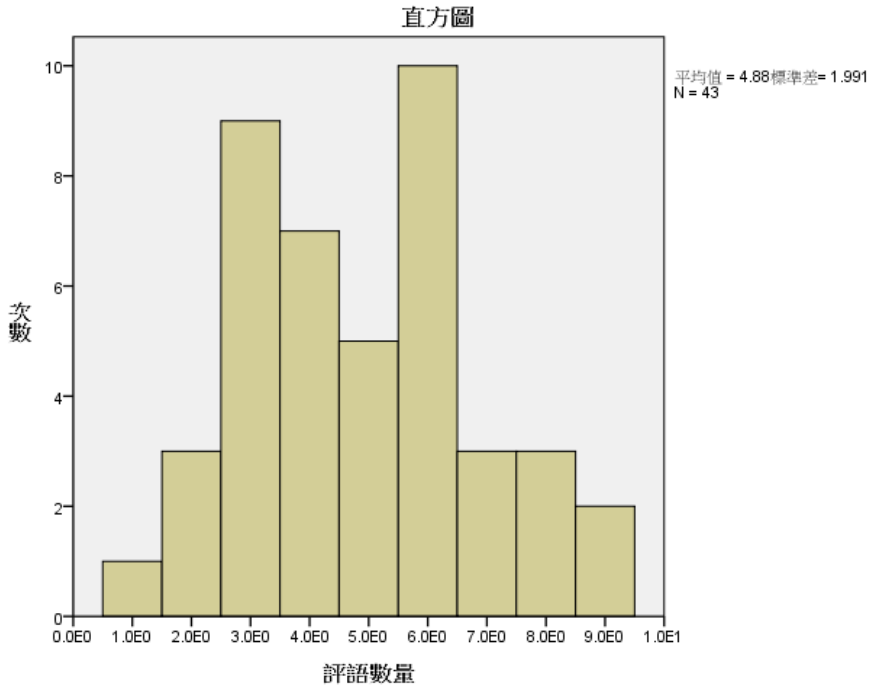
Each leaf: 1 case(s)

doi:10.6342/NTU201602255

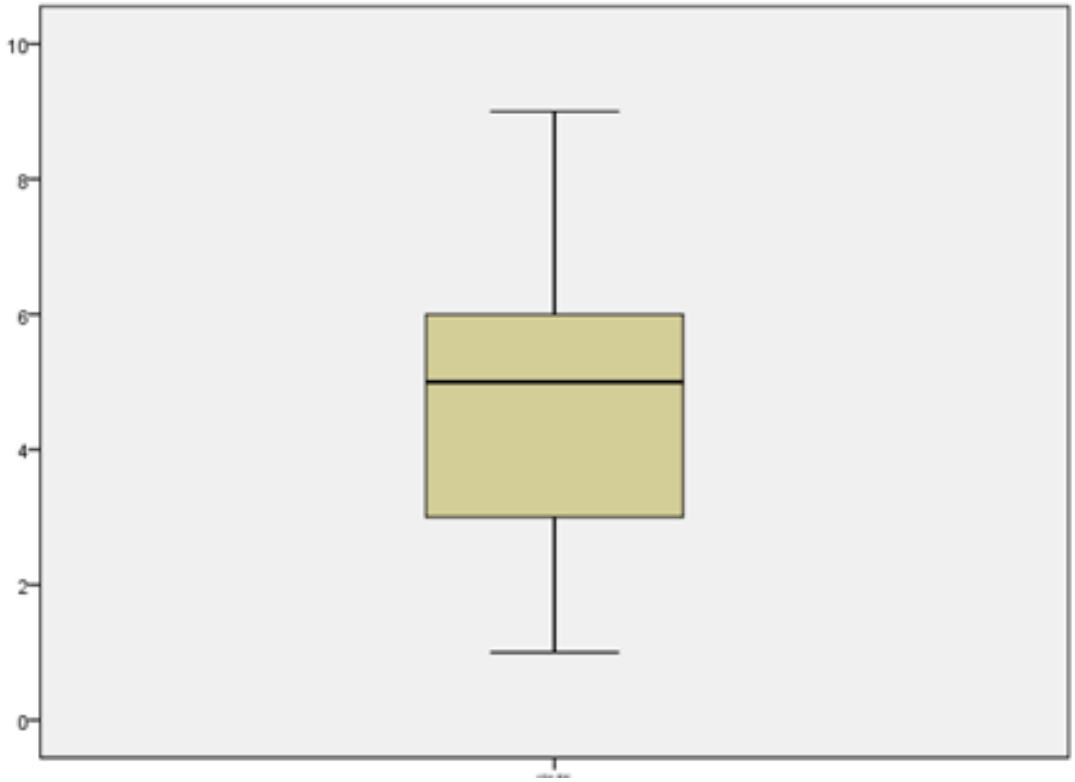
三、量化研究

描述性統計量

			統計量	標準誤
評語數量	平均數		4.88	.304
	平均數的 95% 信賴區間	上限	4.27	
		下限	5.50	
	刪除兩極端各 5% 觀察值之平均數		4.84	
	中位數		5.00	
	變異數		3.962	
	標準差		1.991	
	最小值		1	
	最大值		9	
	範圍		8	
	四分位全距		3	
	偏態		.244	.361
	峰度		-.628	.709



doi:10.6342/NTU201602255



評語數量 Stem-and-Leaf Plot for
量化研究

Frequency	Stem &	Leaf
1.00	1 .	0
3.00	2 .	000
9.00	3 .	000000000
7.00	4 .	0000000
5.00	5 .	00000
10.00	6 .	0000000000
3.00	7 .	000
3.00	8 .	000
2.00	9 .	00

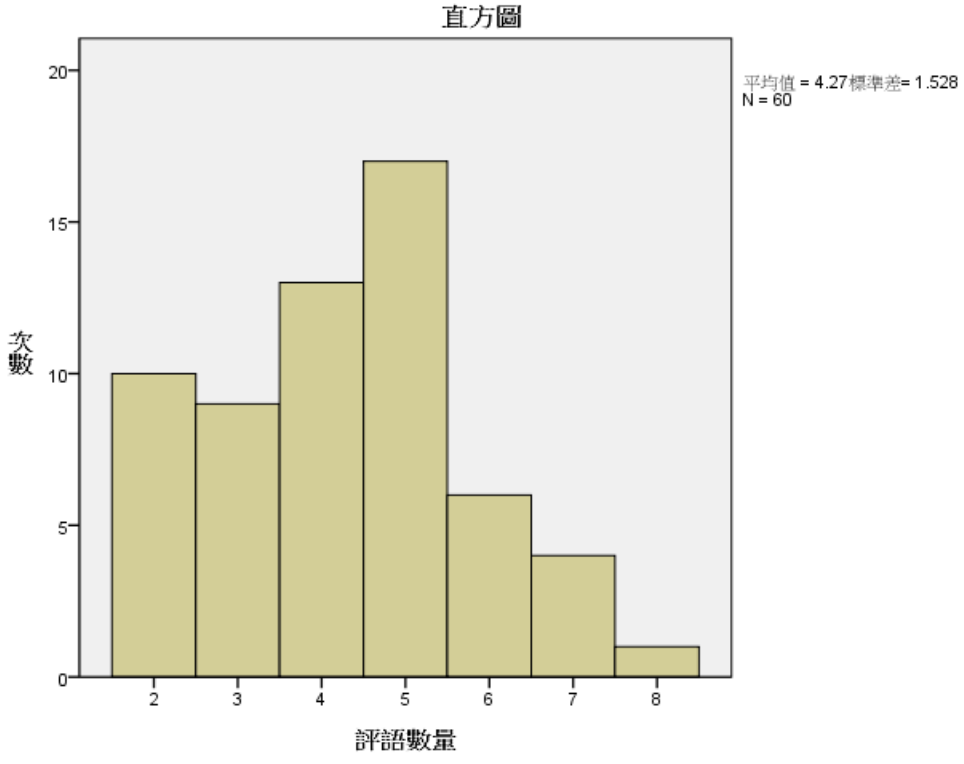
Stem width: 1
Each leaf: 1 case(s)

doi:10.6342/NTU201602255

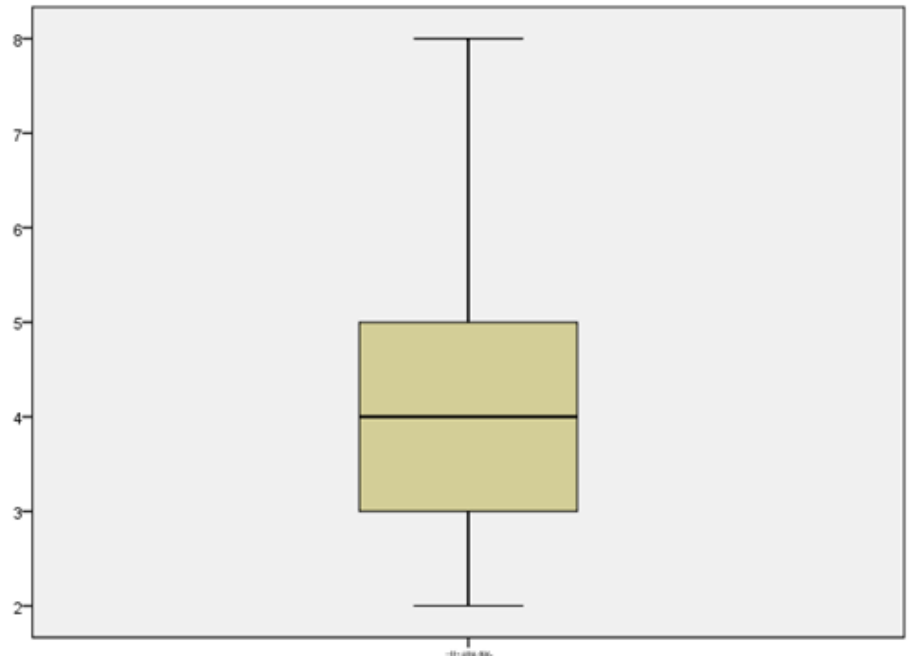
四、非量化研究

描述性統計量

			統計量	標準誤
評語數量	平均數		4.27	.197
	平均數的 95% 信賴區間	上限	3.87	
		下限	4.66	
	刪除兩極端各 5% 觀察值之平均數		4.22	
	中位數		4.00	
	變異數		2.334	
	標準差		1.528	
	最小值		2	
	最大值		8	
	範圍		6	
	四分位全距		2	
	偏態		.180	.309
	峰度		-.529	.608



doi:10.6342/NTU201602255



評語數量 Stem-and-Leaf Plot for
非量化研究

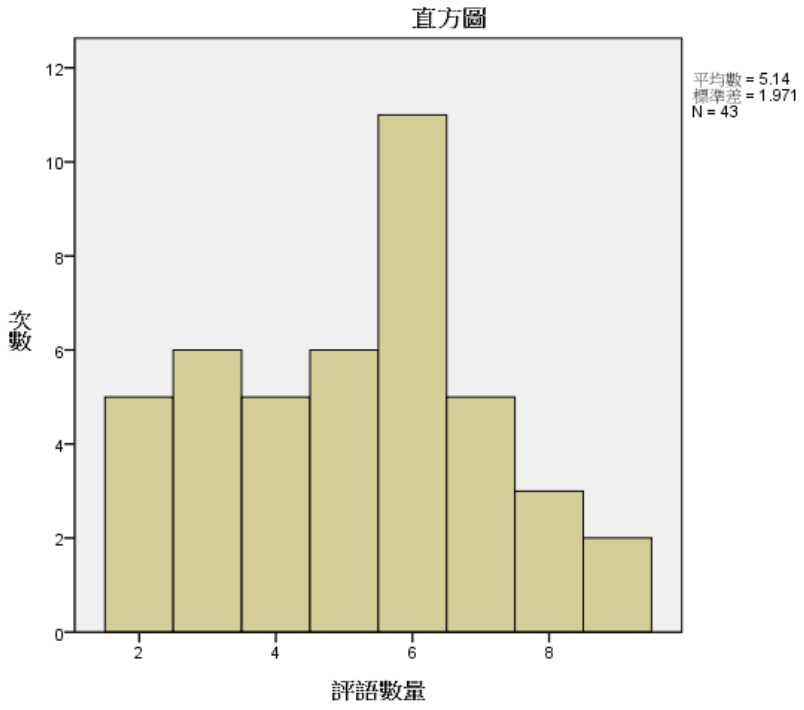
Frequency	Stem &	Leaf
10.00	2 .	0000000000
9.00	3 .	000000000
13.00	4 .	0000000000000
17.00	5 .	00000000000000000
6.00	6 .	000000
4.00	7 .	0000
1.00	8 .	0

Stem width: 1
Each leaf: 1 case(s)

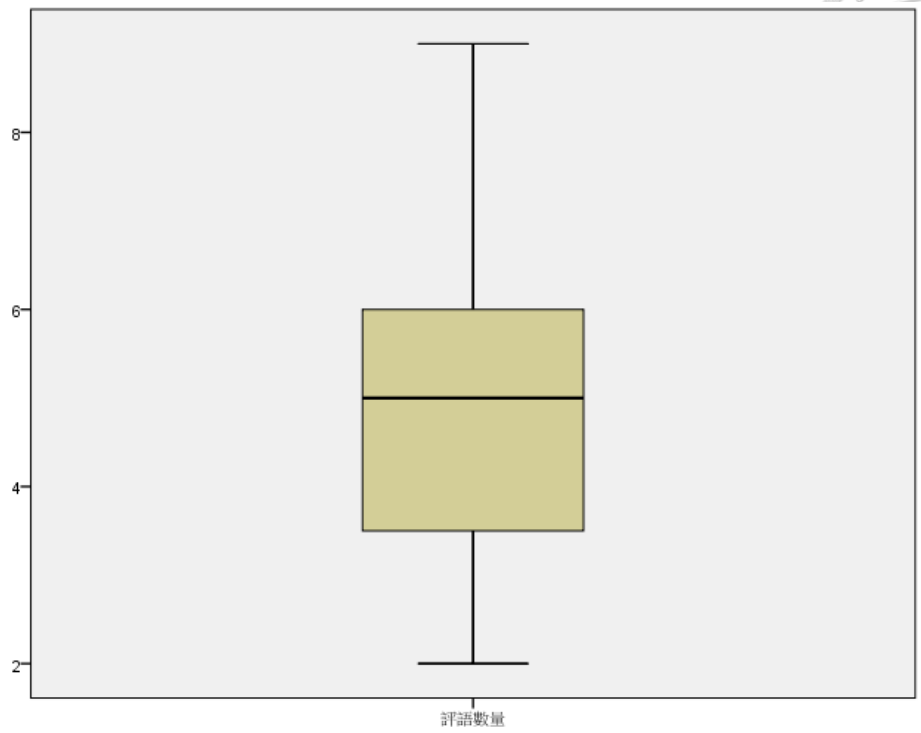
doi:10.6342/NTU201602255

描述性統計量

		統計量	標準誤
評語數量	平均數	5.14	.301
	平均數的 95% 信賴區間 下限	4.53	
	上限	5.75	
	刪除兩極端各 5% 觀察值之平均數	5.10	
	中位數	5.00	
	變異數	3.885	
	標準差	1.971	
	最小值	2	
	最大值	9	
	範圍	7	
	四分位全距	3	
	偏態	.012	.361
	峰度	-.804	.709



doi:10.6342/NTU201602255



評語數量 Stem-and-Leaf Plot

社會科學研究

Frequency	Stem &	Leaf
5.00	2 .	00000
6.00	3 .	000000
5.00	4 .	00000
6.00	5 .	000000
11.00	6 .	00000000000
5.00	7 .	00000
3.00	8 .	000
2.00	9 .	00

Stem width: 1

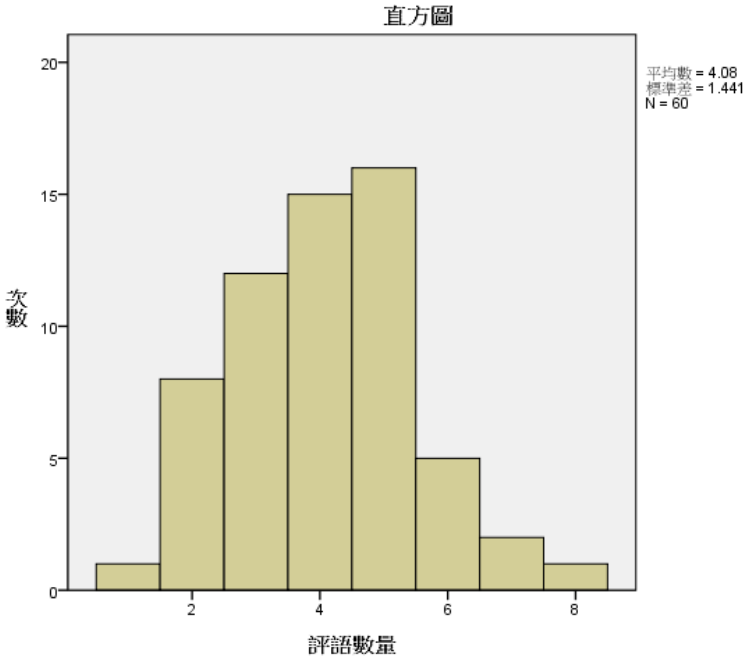
Each leaf: 1 case(s)

doi:10.6342/NTU201602255

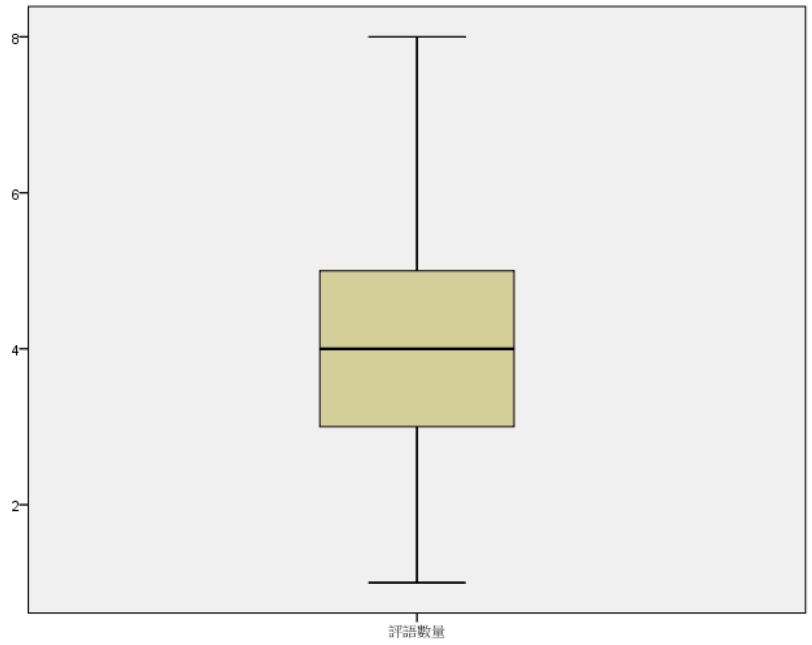
六、人文科學研究

描述性統計量

		統計量	標準誤
評語數量	平均數	4.08	.186
	平均數的 95% 信賴區間 下限	3.71	
	上限	4.46	
	刪除兩極端各 5% 觀察值之平均數	4.04	
	中位數	4.00	
	變異數	2.078	
	標準差	1.441	
	最小值	1	
	最大值	8	
	範圍	7	
	四分位全距	2	
	偏態	.236	.309
	峰度	-.045	.608



doi:10.6342/NTU201602255



評語數量 Stem-and-Leaf Plot

人文科學研究

Frequency	Stem &	Leaf
1.00	1 .	0
8.00	2 .	00000000
12.00	3 .	000000000000
15.00	4 .	000000000000000
16.00	5 .	0000000000000000
5.00	6 .	00000
2.00	7 .	00
1.00	8 .	0

Stem width: 1

Each leaf: 1 case(s)

doi:10.6342/NTU201602255

doi:10.6342/NTU201602255