

論心理測驗與國家考試

陳皎眉* 黃富源** 孫旻暉*** 李睿杰****

摘要

目前國家考試主要以紙筆測驗評鑑應試者的專業知識，但一方面紙筆測驗的效、信度常受到質疑，另一方面，單以紙筆測驗是否能選出最適合特定職位的人選也常受到討論。因此，如何利用良好、多元的測驗方式甄選出最適合特定職務之人選，實為國家考試的核心議題。本文首先以行為樣本、標準化、客觀性概念定義心理測驗。其次說明依心理特徵來分類的不同測驗類型，包括智力測驗、性向測驗、成就測驗、人格測驗等，並探討心理測驗在人才甄選上的兩個重要用途：評鑑應徵者當前特性，與預測未來工作表現。最後討論評估心理測驗良窳的信度和效度，及如何使用心理測驗以挑選最適當的人才。信度指測驗結果的一致性，效度指測驗能夠測量到所欲測量特質的程度，其中效度又比信度更重要。測量效度（內容效度和構念效度）越高越能評鑑應徵者當前特性，決策效度（效標關聯效度）越高越能預測應徵者未來工作表現，因此，國家考試若於甄選過程中同時使用具備測量效度和決策效度的測驗，將更能選拔出適合特定職位的人才。

關鍵字：心理測驗、國家考試、信度、效度、人員甄選

* 考試院考試委員、國立政治大學心理系教授 (Minister without Portfolio, Examination Yuan, R. O. C. & Professor, Department of Psychology, National Chengchi University)

** 考試院考試委員、中央警察大學教授 (Minister without Portfolio, Examination Yuan, R. O. C. & Professor of the Central Police University)

*** 中山醫學大學心理學系助理教授暨附設醫院醫學教育部研究顧問 (Assistant Professor, Department of Psychology, Chung Shan Medical University, R. O. C. & Research Consultant, Department of Medical Education, Chung Shan Medical University Hospital, R. O. C.)

**** 國立政治大學心理系研究生 (Graduate student, Department of Psychology, National Chengchi University)

Psychological Test and National Examination

Jeaw-Mei Chen* Fu-Yuan Huang**
Mein-Woei Suen*** Ruei-Jie Lee****

Abstract

National examinations mainly use written tests to evaluate candidates' professional knowledge. But the reliability and validity of these tests are often questioned. While, whether written tests are able to select the most appropriate candidates is another question. The core issue of national examination, thus, is how to use diverse and valid tests to select the best candidates. This article first defines the psychological tests by behavior samples, standardization, and objectivity. Then, describes the different types of psychological tests with different psychological properties, and discusses the two main applications of psychological tests: evaluate candidate's current characteristics, and predict their future work performance. Finally, the authors discuss how to estimate the reliability and validity in assessing the quality of psychological tests, and on how to use tests in selecting the best candidates. Reliability refers to the consistency of measurement, and validity is the extent to which the test measures what it claims to measure. Validity is more important than reliability. The higher the validity of measurement (content validity or construct validity), the more accurate it evaluates the candidate's current characteristics. The higher the validity of decision (criterion-related validity), the better it can predict candidate's future work performance. Therefore, decision-makers can select the best fit candidates by using tests that possess both validity of measurement and validity of decision.

Keywords: psychological test, national examination, reliability, validity, personnel selection

壹 前言

當前國家考試主要分為公務人員考試及專門職業及技術人員考試，但無論是哪一種考試，其目的皆是為了甄選出最適合其職務之人選。公務人員任用法第二條指出：「公務人員之任用，應本專才、專業、適才、適所之旨，初任與升調並重，為人與事之適切配合」。可見國家考試的基本宗旨除了選出最具有專業能力的人才外，更重要的是要選出最適合特定職位的人才。

公務人員考試法第八條明訂：「公務人員考試，得採筆試、口試、測驗、實地考試、審查著作或發明、審查知能有關學經歷證明等方式行之」，因此國家考試是可以，也建議採用多元的考選方法進行甄選的。然而現今國家考試的形式，有極大部分仍只採用紙筆測驗的形式去評鑑應試者基本知能與專業知識，因此造成了「最會考試的人常常不是最適用的人」。

有鑑於此，考試院（2009）近期所提出之文官制度興革規劃方案，針對「精進考選功能」的議題，提出兩點主要需改進之處：（1）考試方式，亟待多元化（2）應試科目，與職務核心工作職能尚有差距。可見未來國家考試應朝「確立核心職能」、「核心職能與考試方式的配合」、「多元化的考試方式」去努力，也就是先建立各個職務所需之核心職能，並由特定職能發展可評估出此職能之測驗，最後再以多元化的施測方式測量各種核心職能藉以選拔出最適人才。

本文將由心理測驗的角度出發，藉由闡述心理測驗的基本概念、心理測驗的類型與用途、評估心理測驗優劣的準則、和如何利用心理測驗挑選最適人才等，說明心理測驗和國家考試人才甄選的關係。

貳 心理測驗的基本概念

用人主管總是在問，為什麼選不到適合的人才？考試究竟有沒有效？國家考試納入心理測驗就能找到適合的

人才嗎？所謂心理測驗是指對行為樣本（behavior samples）以系統化的步驟加以觀察並將其結果以數字表現出來的

方法(Cronbach, 1984)。另一位學者Anastasi則認為心理測驗是一種對行為樣本所做的一種具標準化和客觀化的測量(Anastasi, & Urbina, 1997/2003)。即使學者對心理測驗的定義稍有不同,但幾乎所有的學者都同意所謂的心理測驗應該有三個共同屬性,分別是行為樣本(behavior samples)、標準化(standardization)和客觀性(objectivity)。

「行為樣本」指在測驗中讓受試回答的問題或操作的作業(task)(郭生玉, 2004)。由於測驗只能測量外在行為,無法測量內在心理特質,因此施測者只能從個人回答測驗問題的行為表現推論其心理屬性。又因為每種心理屬性都包含廣大的行為範圍(behavior domain),而測量所有相關的行為表現又不切實際,所以我們只能從欲測量之心理屬性中,以「抽樣」方式抽取出相關的「行為」樣本,並從受試者在這些樣本的表現推論至特定目標屬性。因此,以國家考試為例,每一考科的測驗試題代表性便是決定測驗品質的重要關鍵,當測驗試題越具有代表性,越能測到欲測量的心理屬性。例如,應試科目:「英文」的考題,如果包含了十個題目,但全部的題目皆為字彙題,這樣的試題便缺乏代表性,無法涵蓋所欲測驗的能力。而如果它有較多類型的題目,包括字彙、文法、聽說能力…則較能測到真正的「英文能力」。

「標準化」指測驗編製、測驗施測

情境與測驗施測步驟均需具一致性。首先,測驗編製的標準化步驟,主要包含測驗架構建立、試題發展與編寫、預試與試題修訂、信效度證據蒐集與檢驗、常模建立、編寫指導手冊或說明書等過程。其次,測驗施測情境的標準化必須考量如光線、溫度、座位安排、安靜程度等。最後,測驗施測步驟的標準化必須考量如指導說明、施測時間、處理試題疑問、試卷發放的順序等。以諮商心理師考試為例,考選部明訂應試科目為人類行為與發展、諮商與心理治療理論、諮商與心理治療實務(包括專業倫理)、團體諮商與心理治療、心理測驗與評量、心理衛生(包括變態心理學),並經由出題委員根據各考科之命題大綱編制試題題庫,再從中挑選適合的試題編制成正式試卷,就是測驗編製標準化的例子。另外,為了維持考場溫度的一致性,目前的考場一律不開冷氣,此為施測情境標準化的一例。最後,無論考場在哪,各個考科皆有固定的測驗時間;而考試完畢後,試題疑義的處理也會有特定會議的召開,這些都是施測步驟標準化的例子。

「客觀性」指在測驗過程中要盡量將人為的主觀因素減至最低,使測驗的結果較不受主觀因素的支配和影響。也就是當兩位評分者針對同一份測驗採用相同的計分規則計分時,相同的作答反應會得到相同(或相似)的計分。為了達到測驗計分的客觀性,就必須量化

受測者在測驗反應行為的計分規則或程序，這些規則或程序必須是夠周延且定義明確，才能使不同的計分者對相同的反應做出相同（或相似）的計分（Murphy, & Davidshofer, 2001/2009）。以牙體技術師的實地考試「全口活動義齒排列」測驗為例，考選部明訂五種評估測驗表現的面向與配分，包含齒列及牙弓的對稱性、前牙的水平覆蓋及垂直覆蓋關係、後牙的一齒對二齒的咬合關係、補償曲線的付予、牙肉形成含齒頸的等高線及對稱性等，且每一面向的配分各為 20 分，總分為 100 分，此外若是不依規定使用統一模型及人工齒排列者，或是人工齒經由修整而致外觀改變者，或是牙齒部位上下左右前後相反者，皆不予計分。

由上述對心理測驗的定義可知，不同形式的國家考試基本上都可算是一種心理測驗，因為它們都具備了行為樣本、標準化、客觀性等三項心理測驗的

基本屬性。以筆試的測驗形式為例，高等考試中行政法測驗的每一項試題，皆是為了測量受測者對行政法瞭解程度的行為樣本；再以口試的測驗形式為例，高等考試第二級的口試測驗中，口試委員所提出的問題如：「談談最近一次你所運用的最有創意的解決問題方法是什麼？」即為一種行為樣本；最後以實地測驗的測驗形式為例，鐵路人員養路工程類科的體能測驗，是以「能在二十秒內負重四十公斤沙包直線走跑四十公尺」為體能測驗的行為樣本；除了行為樣本之外，在不同的測驗形式，如上述所提的筆試、口試、實地測驗等，考選部也會於每次考試前依據標準的測驗編製程序編製當次測驗的試題（標準化），並於出題完畢後，明訂各項試題的計分規則及程序，使閱卷（或計分）人員可依客觀的計分規則及程序計算出總分（客觀性）。

參

心理測驗的類型與用途

一、心理測驗的類型

心理測驗涵蓋的範圍極廣，種類也相當繁多，而依據分類方式的不同，又可分成多種不同類型的測驗（郭生玉，2004），譬如依據心理特徵來區分，可分為智力、性向、成就和人格測驗；

依據測驗的材料，可分為文字和非文字測驗；依據測驗的人數，可分為個別測驗和團體測驗；依據測驗時限，可分為速度測驗（在有限時間內，測量完成試題的速度）和難度測驗（在充裕時間內，測量解決問題的能力）；依據測驗

的反應形態，可分為最大表現測驗（測量個人的最佳反應、最大成就）和典型表現測驗（測量個人通常所表現出的行為）；根據測驗的解釋方式，可分為常模參照測驗（根據測驗分數在團體中的相對地位加以解釋的測驗）和效標參照測驗（根據測驗分數在施測前所訂定的標準加以解釋的測驗），例如專技人員高等考試的律師考試即為一種常模效標測驗，當應考人總成績平均分數達全程到考人數的前百分之八，才達錄取標準。又例如專技人員高等考試的營養師考試即為一種效標參照測驗，當應考人總成績平均滿六十分即達及格標準。上述對心理測驗的分類方式，又以依心理特徵來區分的方式最重要且廣為人知，故以下針對這四類測驗作簡單介紹。

智力測驗（intelligence test）是用來測量個人的一般能力，包含記憶、推理、思考、歸納、演繹、想像、創造、評鑑等等能力，又稱普通能力測驗（general ability test）。

性向測驗（aptitude test）是在測量學習的潛在能力，亦即測量個人未來發展的可能性。通常性向測驗又分普通性向與特殊性向，普通性向即學習一般事物所需的共通能力，而特殊性向即學習某些特殊才能所需的能力，如音樂、美術、機械等特殊才能。

成就測驗（achievement test）是在測量由教育或訓練所獲得的實際能力。成就測驗又可依功用和內涵分為綜合成

就測驗、特殊成就測驗、與診斷測驗三類。綜合成就測驗旨在一個測驗中測量不同學科的成就水準；特殊成就測驗在測量某一學科的成就水準；而診斷測驗則在測量學習的困難所在，以便於進行教學或輔導所需。

人格測驗是在測量個體的各種人格特徵，包含態度、動機、興趣、價值觀、自我概念、情緒、人際關係、人格特質等。人格測驗通常又可依受測者所進行測驗作業之結構和清晰程度，分為客觀測驗（objective test）和投射測驗（projective test）。客觀測驗通常具高度結構性，能客觀計分，測驗題目也較明確且清晰；投射測驗則較不具結構性，使用模糊不清的測驗題目（例如圖片或墨漬），讓受測者依自己潛藏的期望、態度及需求來作答，再由其回答推論其內在人格特質（Murphy, & Davidshofer, 2001/2009）。

國家考試中，公務人員和專技人員考試的筆試大都是一種成就測驗，每一考科的測驗目的都在於測量受測者對於該科目的熟練程度。至於部分國家考試所採用的口試則較偏像是一種特殊性向測驗，希望能用以測出受測者在口語表達能力、社交能力、危機應變能力等特殊才能。

二、心理測驗的用途

心理測驗的主要功用在於協助個體做出重要的決策（Murphy, & David-

shofer, 2001/2009)。但因為心理測驗所欲測量的心理屬性經常受生理、心理和社會等其他多重因素影響，因此心理測量常無法像物理測量那樣單純且直接。即便如此，美國國家科學院（National Academy of Sciences）仍指出：心理測驗是最佳、最公平、且最經濟來獲取有用資訊的方法，以協助個人做出合理的決策（Psychological tests generally represent the best, fairest, and most economical method of obtaining the information necessary to make sensible decision about individuals）（Wigdor & Garner, 1982）。因此，雖然心理測驗本身並非完美，但當個體欲做出重要決策時，它仍是最佳、最公平，也是最精確的一項技術（Murphy, & Davidshofer, 2001/2009）。

目前廣泛使用心理測驗協助進行決策者，主要可分為教育、人事及臨床三大領域。在教育領域中，心理測驗最常被用來進行入學考試及課堂考試，前者如每年進行的國中基本學力測驗和大學指定科目考試，後者如學校各學科之期中、期末或隨堂考試，都是心理測驗在教育領域應用的例子。在人事領域中，心理測驗經常被用來作為人事甄選的工具，例如每年舉行的高、普考便是政府利用測驗做為甄選人才的方式。在臨床領域中，心理測驗也被利用來做為臨床衡鑑的工具之一，例如明尼蘇達多項人格量表（Minnesota Multiphasic Personality Inventory, MMPI）、羅夏克墨漬測驗

（Rorschach Inkblot Test）等量表。

在政府和企業中，心理測驗最常被運用來做為人事領域的用途，其中又以在甄選過程中最常被使用。所謂甄選就是組織就其所設的職位，蒐集並評估有關應徵者的各種資訊，以作出聘僱決定（吳復新，2003）。甄選的目的有二，一為評鑑（assessment），另一為預測（prediction）。評鑑意指客觀評估求職者所具之特性，以決定其合乎工作要求的程度；預測則是由求職者目前所具有的特性（包括表現出來的行為）預測其將來的工作績效。故若能使用有效的測量方法，確實辨認出應徵者所具備的工作知識、能力與特質等條件，同時又確知這些條件與工作績效的關係，便能做好甄選的工作（吳復新，2003）。

在企業甄選中，目前已大量採用多種心理測驗做為決策輔助工具，包括認知能力測驗、人格測驗、面談、評量中心（assessment centers）等，以求達到評鑑求職者當前特性和預測未來工作績效兩種甄選目的。研究結果也發現，上述心理測驗在預測求職者的後續工作績效上，常具有不錯的水準（Hough, & Oswald, 2000; Salgado, 1999）。至於政府的人才甄選方法中，目前仍以測量工作知識的紙筆測驗為主要甄選方法，部分考試則加入了口試、實地測驗等不同的測驗形式。然而，這些測驗多半是以評估求職者的當前特性與能力為主要目的，對於預測求職者未來工作績效上可

能不是那麼有幫助。因此，未來國家考試的甄選方法，應加入更具備預測效用

的測驗形式，才能達到人才甄選在預測工作績效上的目的。

肆

評估心理測驗優劣的準則

早在古代的中國就以科舉考試做為拔擢優秀人才的方法，直到現今，國家公務人員、專技人才的選拔仍以考試為主要甄選方式。因此，測驗本身品質的優劣便是影響能否選到優秀且適任人才的關鍵，如果測驗的品質良好，據之所做的決定當然會比品質較差的測驗來得更好。那麼，我們應依據什麼標準來判斷一個測驗的品質優劣？一般來說，測驗的信度（reliability）與效度（validity）是最常被用來評估測驗優劣的準則，以下將分別介紹之。

一、信度

信度是指測驗結果的一致性（Reliability refers to the attribute of consistency in measurement）（Gregory, 2007）。更詳細來說，信度代表相同的個人在不同的時間，以相同的測驗測量，或以複本測驗（相等的試題）測量，或在不同的情境下測量，所得結果的一致性（Anastasi, 1988）。一般而言，心理測驗的信度較自然科學測驗來得低，因為心理測驗所測量的心理特徵並無法直接測量，而只能透過間接測量而推論之。

如同測量體重一般，你不會在每次測量時都得到一模一樣的體重數字，因為體重數字經常會受許多其他因素所干擾，而導致最後結果的不同。信度的計算也是基於相同的道理，心理測驗的測量結果同樣也會受許多因素所影響，因此在考慮測量結果的穩定性時，也必須考慮可能的測量誤差有哪些，當測量誤差越小時，測量結果也就越穩定一致，此時測驗便具有較高的信度。

由於在心理測驗中誤差來源相當多元，測驗編製者無法完全將它們一一排除，因此必須以統計分析的方式去估計信度高低，信度的估計與測驗的誤差兩者密不可分。測驗的誤差來源主要來自：1. 測驗編製（test construction），如題目取樣、測驗內容取樣等；2. 測驗實施（test administration），如測驗情境（溫度、燈光、通風、安靜程度等）、受試者（動機、注意力、情緒等）、主試者（外表、態度、指導說明等）；3. 測驗計分（test scoring），如申論題、作文之計分、個別智力測驗、創造力測驗、口試測驗等主觀計分方式（郭生玉，2004）。常用的信度估計方法可依據測

驗的誤差來源做分類，內部一致性信度（internal consistency reliability）和測驗編製的誤差有關，再測信度（test-retest reliability）和測驗實施的誤差有關，複本信度（alternate-forms reliability）與測驗編製、測驗實施的誤差皆有關，評分者信度（scorer reliability）和測驗計分的誤差有關。

常見的內部一致性信度估計法可分為折半信度（split-half reliability）、庫李信度（Kuder-Richardson reliability）、 α 係數（alpha coefficient）、和複本信度。基本上這些信度指標都是以測驗題目內容的一致性為基礎，題目的同質性越高，分數的意義越明確，信度也越高。當測驗編製時，各個測驗題目所測量的性質必須越相近，否則便是一份編製不良的測驗，也不具有內部一致性信度。

再測信度是指使用同一個測驗在不同的時間，重複測量相同的一群受試者，根據這兩次分數求得的相關係數即為再測信度。當此相關係數越高，代表兩者的一致性越高。然而，再測信度的缺點在於易受練習與記憶影響，為了降低練習與記憶殘留對信度估計的影響，複本信度於是更常被採用。複本信度指在測驗發展之初便編製兩份內容相近的測驗，之後請受測者先做一個版本，經過一段時間後再做另一個版本，受測者在兩份測驗上得分的相關係數即為複本信度。無論是再測信度或複本信度皆會進行兩次的施測，並計算出測驗實施所

帶來的誤差，故可被歸類於和測驗實施誤差有關的信度估計法。

由於再測信度與複本信度的估計均需實施兩次測驗，有較多外在因素影響測驗的實施，相對於只需從一次的測驗結果即可估計信度的內部一致性信度估計方法，不僅減少編製測驗所耗費的時間精力，也較方便實施，故由單次測驗估計信度的內部一致性信度相較於再測或複本信度，更常被用來計算信度。

評分者間信度是指針對相同的試卷答案，由兩位（或多位）評分者分別評分所得分數之相關係數。評分者信度主要用於主觀計分之測驗，如申論題式的成就測驗、投射測驗、創造力測驗、口試測驗等，其評分會受到不同評分者主觀判斷影響的測驗。因為相較於有明確訂定計分規則的客觀計分方式，採用主觀計分的測驗會受到評分者主觀判斷的影響，導致計分的誤差，此時評分者信度便可做為評斷此種主觀計分方式所得分數的一致性程度。

一份測驗的信度，亦即測量結果的一致性或穩定性是判斷測驗是否有效的先決條件，若每次的測量結果都不同或是測驗內容分歧，我們將無法確認這份測驗到底在測量什麼。故一份測驗要能有效測量目標特質，具備一定程度的信度是必要的條件。然而，一份測驗具有高信度，並不保證也會有高效度，即便每次測驗所得的分數都很一致，但若測驗所測量到的特質不是目標特質，此份

測驗也不具效度。舉例來說，一個體重計每次測量都少了五公斤，表示測量結果很一致（有信度），但卻不是正確的體重（無效度）。因此我們可以說，信度與效度是評斷測驗優劣的兩個最主要指標，其中效度又比信度重要，它是測驗工具優劣的最重要指標（Ahmann & Glock, 1981; Mehrens & Lehmann, 1991; Nitko, 1996）。

下圖可以說明信度與效度的關係，當射箭很有信度時，如左邊或中間的標靶，箭頭很一致集中在某處；當射箭很有效度時，如左邊的標靶，箭頭一致的集中在靶心處。由此可知效度需要同時具備一致性和正確性，而信度只需一致性即可。

二、效度

效度是指一份測驗能夠測量到所欲測量特質的程度（The validity of a test

is the extent to which it measures what it claims to measure）（Gregory, 2007）。一份測驗越能真實且正確的測量目標特質，這份測驗的效度就越高。基本上，效度的特徵在於它無法直接被測量，但可從其他資料推論之。此外，效度只有程度上的差別，並非全有或全無，也就是說沒有一個測驗具有完全的效度，也沒有一個測驗完全沒有效度，它只是程度上的不同。最後，效度在使用目的和情境方面具有特殊性，可能對某一測驗目的而言，此測驗有很高的效度，但對另一個目的而言，就可能幾乎不具任何效度。譬如一份英文閱讀測驗能有效的測量出個體的英文閱讀能力，但卻完全無法測量到他的英文口語表達能力。因此，選擇測驗時需依據使用的目的挑選具適當效度的測驗。

一般來說，心理測驗的效度可依照其用途分為兩種效度形式，一種為

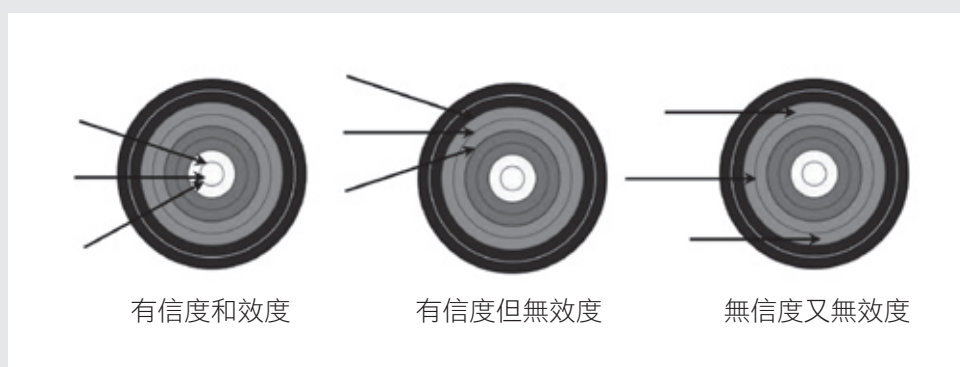


圖1 以射箭說明信度與效度關係（修改自郭生玉，2004）

測量效度 (validity of measurement)，代表此測驗具有能測量個人某種屬性的功用；另一種為決策效度 (validity of decision)，代表此測驗具有預測未來的結果並協助個人做出有關決策的功用 (Murphy, & Davidshofer, 2001/2009)。這兩種效度也反應人事甄選的兩個主要目的，測量效度可用來協助評鑑求職者的當前特性，決策效度則可用來協助預測求職者未來的工作績效。

由於效度無法直接被測量，測驗編製者必須從各種不同的來源蒐集相關證據，以證明測驗的確是在測量它原先被設計用來測量的屬性。其中在蒐集測量效度證據時主要可從兩種不同的角度來區分，分別是內容導向的驗證策略和構念導向的驗證策略，由內容導向所得出的效度稱為內容效度 (content validity)，由構念導向所得出的效度稱為構念效度 (construct validity)。此兩種驗證策略都是在累積那些指出測驗確實在測量原本所欲測量屬性的證據。內容效度是透過測驗內容本身來建立，構念效度則是透過測驗分數與其他多種不同測量方法間的關係來建立。

內容效度是測驗能否適切地針對內容領域進行抽樣的判斷 (Guion, 1977)，也就是指測驗內容的代表性或取樣的適切性，或是測驗題目是否是欲測量屬性的代表性樣本。內容效度的驗證主要採邏輯的分析方法，首先可判斷題目的適切性，即題目是否和所欲測量

的內容或能力有關聯；接下來判斷測驗內容的代表性，即測驗中的題目是否能涵蓋所欲測量的內容領域。若是題目和欲測量屬性間有高度關聯，且涵蓋的範圍越周延，則內容效度越高。例如臨床心理師考試的臨床心理學總論（一）此一考科，主要在測量應考人對「偏差行為的定義與描述」、及「偏差行為的成因」兩大主題的理解程度。假如試卷中的所有試題都在測量「偏差行為的定義與描述」這部分的內容，包括偏差行為的定義、描述、分類、研究方法等，此時即使這些考題與「臨床心理學（一）」有高度關連性，卻缺乏測量「偏差行為的成因」的試題，因此這份試卷的內容效度並不高。而當試卷中的試題同時包含了「偏差行為的定義與描述」及「偏差行為的成因」時，此份問卷才具有和欲測量內容領域間的關聯性與代表性，也就是具備高內容效度。

構念效度是指測驗能夠測量或說明理論上的構念、心理學概念、或特質的程度 (Anastasi, 1988)。構念是心理學上的一種理論構想或特質，它無法被觀察到，卻被假設是存在的，以便解釋個人的行為，例如智力、動機、人格等。這些構念都有其理論基礎，並可依據理論預測個體行為。由於構念無法被觀察到，因此要確定一個測驗是否提供良好的題目代表其所測量的構念，就必須要將抽象的構念轉換為具體可觀察的行為。當有越多的相關行為說明構念的意

義，構念的定義就越清楚明確。譬如五大人格理論（five factor theory）中的外向性（extraversion）特質便是一種心理構念，我們無法觀察到它，只能透過相關的行為描述，例如「會主動與人交談」、「總是熱情和他人打招呼」等來描述外向性這個特質，當有越多相關行為來描述外向性時，我們對外向性構念的瞭解就越清楚。

考試院正在研擬用來做為甄選用途的文官人格特質量表，主要整合三個不同的角色特質，也就是從一般人、工作者、及文官特質的角度出發，發展合適的文官人格特質量表（陳皎眉、胡悅倫、洪光宗，2010）。「一般人」的特質以五大人格理論中的五項人格特質，包括神經質性（neuroticism）、外向性（extroversion）、開放性（openness to new experience）、和善性（agreeableness）與嚴謹性（conscientiousness）為主。「工作者」以問題解決、人際溝通、決斷力為重要測量特質。「文官」則以廉正、忠誠、專業、效能、關懷為核心特質。最終的文官人格特質量表可能包含許多不同的特質構念，但每一個特質構念均需具備一定程度的構念效度，此量表才可有效應用於甄選用途上。

構念效度並無法用單一統計量數來表示，而是必須從許多資料來源中，逐步累積證據來做綜合判斷。一般而言，確定構念效度的方法相當多樣，

包括相關研究（correlation study）、團體差異的分析（group differences）、實驗研究（experimental study）、內部一致性分析（internal consistency）、因素分析（factor analysis）、多項特質－多項方法分析（multitrait-multimethod approach）等（郭生玉，2004）。每一種方法所獲得的資料或統計量都有助於建構效度的判斷，而最佳的驗證方法通常會視想要測量的建構而定。

內容效度與建構效度兩者都在判定測量效度的好壞，也就是判定測驗是否在測量它該測量的。即便如此，某些情況下可能只能從兩種效度擇一做為測量效度，當某個領域的界限與結構較明確時，如統計學的知識，但我們卻不知道統計學的知識如何連結到其他可觀察的特質或行為上，此時內容效度適用來評估測量效度的最可行方法。反之，若我們比較瞭解某個構念與一些可觀察行為間的關係，如「誠實」，卻不知該構念的界限或範圍在哪，此時估計建構效度便較為合適。

國家考試的紙筆測驗，每個考科通常具有明確的範圍或界限，因此通常會以內容效度做為效度好壞的判斷依據。然而，即便每個考科都具有一定程度的內容效度，可以準確測量出應試者對該考科的理解程度，但並不代表此測驗就具有決策效度，可以用來預測求職者未來相關的行為。組織在甄選人員時，預測求職者未來的工作績效通常較評鑑

其當前特性更為重要。但一個具備測量效度的測驗並不代表其決策效度也會很好。不同的測驗有不同的適用範圍，此

時便需要利用其他更具決策效度的測驗，以協助預測求職者未來的行為表現。

伍 如何利用心理測驗挑選最適人才

一、效標關聯效度

決策效度又稱效標關聯效度 (criterion-related validity)，指測驗分數和外在效標之間的關聯程度。所謂外在效標 (external criteria) 即測驗所要預測的某些行為或表現，例如利用智力測驗成績預測工作績效，此時工作績效即為一種效標。如果智力測驗與工作績效之間有高相關，代表它具有良好的效標關聯效度。

效標關聯效度依效標取得時間的不同，分為預測效度 (predictive validity) 與同時效度 (concurrent validity)。預測效度是指測驗分數與測驗實施後一段時間所取得效標資料之間的相關；同時效度指測驗分數與實施測驗同一時間所取得效標資料之間的相關。預測效度被認為是評估效度最精確的方法，因為它直接測量所有申請者的測驗分數（在決策之前即可獲得）和效標表現（在決策之後才能獲得）之間的關聯。然而，這種方法卻有現實上的矛盾，因為要計算最正確的預測效度時，必須對所有人都做出相同決策（例如錄取所有申請者），

不然就是隨機錄取，但實際情況卻經常無法允許此種情況發生。較常用來計算預測效度的方法是以測驗分數做為決定錄取的依據，再以被錄取者之後的效標表現計算預測效度，此種預測效度代表從被錄取者群中所計算出的效度，而非由全部申請者所得之效度。因此，預測效度在實務中常常無法被完全正確的計算出來。

現實情況下，同時效度較常被做為效標關聯效度的估計方法，主要在於它具有兩項優點：第一，同時效度的研究較實用，因為研究者並不需要等待一段時間才取得效標分數。第二，雖然測驗理論指稱同時效度係數嚴重低估母群體效度，但其實同時效度的大小和預測效度類似 (Barrett, Phillips, & Alexander, 1981; Schmitt, Gooding, Noe, & Kirsch, 1984)。因此，同時效度的驗證取向較預測效度的驗證取向更常被做為效標關聯效度的估計法。

二、效度與人員的挑選

得到測驗的效度係數，尤其是效標

關聯效度係數後，下一步便是利用此效度係數進行決策，也就是如何利用測驗挑選出將來工作會成功的人選，並拒絕將來工作會失敗的人選。在此之前，先必須瞭解決斷分數(cutoff score)、錯誤接受與錯誤拒絕(false acceptance & false rejection)、基礎率(base rate)、挑選率(selection ratio)等概念。

決斷分數是決定錄取與否的分數，當應徵者的測驗分數高於決斷分數時，即被錄取為正式員工。決斷分數又與錯誤接受、錯誤拒絕兩種選擇的誤差有關，若決斷分數定的很低，則會有許多錯誤接受失敗應徵者的情況產生；若決斷分數定得很高，則會有許多錯誤拒絕成功應徵者的情況產生。較好的決斷分數應盡可能選出工作成功者，也就是正確接受未來會成功者，並排除工作失敗者，也就是正確拒絕未來會失敗者為佳。譬如專技高考會計師考試中的專業科目是以考試分數達60分為及格，應考人需在全部的專業科目考試皆拿60分以上才可取得會計師執照，此時60分即為決斷分數。當應考人分數未達60分而未取得會計師執照，但如果其實他具有在後續工作表現良好的能力，此時便是做了錯誤拒絕的決策；若應考人分數達60分並取得會計師執照，但其實他並無法勝任後續的工作時，此時便是做了錯誤接受的決策。

基礎率是指在應徵者母群體中，工作上會成功者的百分比。假如有95%

的應徵者在工作中都能成功，那基礎率即為.95，若只有5%的應徵者能成功，則基礎率為.05。故基礎率可顯示母群體中有多少百分比的人是潛在成功者。一般而言，基礎率越接近.50，測驗的預測越正確，基礎率過高或過低時預測的正確性都會越低(Aiken, 1997)。

挑選率是指應徵者被選擇出來的百分比。例如要在100名應徵者中接受95名，則挑選率為.95，若只接受5名，則挑選率為.05。通常挑選率越高，測驗效度越不具太大參考價值，因為既然要接受這麼多人，則用什麼挑選策略皆影響不大，但挑選率越低時，測驗效度的參考價值就越高，因為有相當多人競爭稀少的空缺，此時一個好的測驗決策效度將有助於挑選到最適合的人選。一般來說，「當挑選率足夠低時，即便是中等效度的測驗，仍可以顯著提昇決策的準確性」(Murphy, & Davidshofer, 2001/2009)。有關基礎率、挑選率與測驗效度如何影響決策精準性，可參考Taylor與Russell(1939)出版的一系列表格。下表1為假定基礎率為.50，也就是預期有50%的應徵者為潛在成功者時，挑選率與效度如何影響預期應徵者將來工作成功的比例。

因此單純只看效度的高低並無太大意義，因為當基礎率或挑選率很高時，即便有效度係數.70的測驗，可能也沒多大用處；反之若基礎率適中且挑選率很低時，即使是效度係數.20的測驗，

表1 基礎率為 .50 時，顯示其預期成功比例的Taylor-Russell 表

效度	挑選率									
	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.10	.54	.54	.53	.52	.52	.51	.51	.51	.51	.50
.20	.67	.64	.61	.59	.58	.56	.55	.53	.53	.52
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.45
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55
.90	1.00	1.00	.99	.97	.92	.86	.75	.70	.62	.56
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56

(修改自 Taylor, & Russell, 1939)

也能大幅提昇決策的品質。因此除非我們瞭解測驗的使用情境，否則無法完全瞭解效度係數的意義。

國家考試的測驗情境，基本上挑選率極低（常小於 .10），但在假設基礎率為 .50（應考人中有一半的人如果給他機會將會成功）的情況下，只要能實施具備中等效標關聯效度（例如 .50 以上）的測驗，就可大幅提昇正確接受工作成功人選的比率（效度 .50，挑選的

人有 88% 會成功，效度 .60，挑選的人有 94% 會成功，效度 .70，挑選的人有 98% 會成功）（Taylor-Russell 表，引自 Murphy, & Davidshofer, 2001/2009）。上述所談的效度不僅適用於能力測驗的效標關聯效度，人格測驗的效標關聯效度也同樣適用，凡是能預測應徵者未來工作表現的測驗，皆可利用於人才挑選上。因此，增加具備效標關聯效度的測驗，應該是未來國家考試的改革目標。

陸 結論

本文藉由心理測驗的基本概念、心理測驗的類型與用途、評估心理測驗優劣的準則、和如何利用心理測驗挑選最適人才等主題，說明心理測驗和國家考試的密切關聯性。測驗對於國家考試的重要性，主要在於協助決策者利用測驗分數，挑選出將來適合從事該職位的人選。然現今以筆試為主的國家考試形

式，常只具備某種程度的測量效度，而不具備可預測求職者未來工作表現的預測效度，故才会有「考試方式，亟待多元化」的改革方針出現。本文在此以簡介心理測驗的各類重要概念做為開端，期望未來國家考試能具備選拔「最適人才」之功用。

參考文獻

- 考試院 (2009)。文官制度興革規劃方案。台北市：考試院。
- 吳復新 (2003)。人力資源管理—理論分析與實際應用。台北：華泰文化。
- 郭生玉 (2004)。教育測驗與評量。台北：精華書局。
- 陳皎眉、胡悅倫、洪光宗 (2010)。人格測驗在國家選才上之使用與發展。國家菁英季刊，第六卷，第四期，1-12 頁。
- Ahmann, J. S., & Glock, M. D. (1981). *Evaluation Student Progress: Principles of Tests and Measurements*. (6th ed.). Boston: Allyn and Bacon.
- Aiken, L. R. (1997). *Psychological Testing and Assessment*. (9th ed.). Boston: Allyn and Bacon.
- Anastasi, A. & Urbina, S. (2003)。心理測驗 (危芷芬譯)。台北市：雙葉書廊。(原著出版年：1997)
- Anastasi, A. (1988). *Psychological Testing*. (6th ed.). New York: Macmillan Publishing.
- Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 66, 1-6.
- Cronbach, L. J. (1984). *Essentials of Psychological Testing*. (4th ed.). New York: Harper & Row.
- Gregory, R. J. (2007). *Psychological testing: history, principles, and applications*. (5th ed.). Boston: Pearson Education, Inc.
- Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, 1, 1-10.
- Hough, L. M., Oswald, F. L. (2000). Personnel selection: Looking toward the future-remembering the past. *Annual Review of Psychology*, 51, 631-664.
- Mehrens, W. A., & Lehmann, I. K. (1991). *Measurement and Education in Education and Psychology* (4th ed.). New York: Holt, Rinehart and Winston.
- Murphy, K. R., & Davidshofer, C. O. (2009)。心理測驗 (王文中、陳承德譯)。台北市：雙葉書廊。(原著出版年：2001)
- Nitko, A. J. (1996). *Educational Assessment of Students*. (2nd ed.). NJ: Prentice-Hall.
- Salgado, J.F. (1999). Personnel selection methods. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial & Organizational psychology*. New York: Wiley.
- Schmitt, N., Gooding, R.Z., Noe, R.A. & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407-422.
- Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology*, 23, 565-578.
- Wigdor, A. K., & Garner, W. R. (1982). *Ability Testing: Uses, Consequences, and Controversies*. Washington, DC: National Academy Press.