國立臺東師範學院特殊教育學系、特殊教育中心
特殊教育學術研討會論文集，民 92，1-18 頁

# 專題演講：
# 特殊教育研究的未來趨勢--
# 以量化研究分析方法為例

吳 裕 益
國立高雄師範大學

## 一、特殊教育量化研究的問題及新趨勢

特殊教育研究的對象較為特殊，長久以來一直面臨數個不易克服的問題。其中較為重要的有以下各項：

(一)特殊學生之定義不明確，特別是學障之定義更是分歧。

由於各類特殊學生之定義模糊，不但很難找到適當的研究對象，而且即使勉強進行研究，每個研究所謂的學障或其他特殊學生幾乎都不相同，當然所得結論也就不具說服力。

有關學障之定義及鑑定問題，吳裕益(民88)在「從測驗與統計原理評現行學習障礙鑑定方式與標準」一文中曾有詳細之討論。

(二)研究樣本取得困難。

特殊教育研究的對象屬於多變項常態分配相當極端的特定群體，這些特殊群體的母群本來就較小，特別是感官方面障礙的特殊學生更少，研究樣本之取得遠較一般教育研究困難，因此特殊教育研究的樣本通常少於一般教育研究，有不少是屬於小樣本的縱貫研究，甚至是單一個案時間系列研究。

小樣本的研究在統計考驗時很難達顯著水準，也就難以將研究結果推論至樣本所代表之母群。目前常用的統計套軟體，所使用的推論統計分析方法是假定母群是無線大的，如果是有限母群的推論，其抽樣誤差較小，需加以調整。另外，如果是單一個案時間系列研究，一般常用的推論統計分析方法也不太適用，需使用適用於單一個案時間系列研究的統計分析方法，如，評鑑「介入效果」的簡易時間系列分析(C 統計數)及單一受試資料實驗效果量(effect size)之分析方法等。另外，如果是多群體的縱貫資料，也可以使用階層線性模式(hierarchical linear model，簡稱 HLM)來分析不同群體成長軌線(截距及斜率)之差異情形，以及影響成長軌線之可能因素。

(三)研究及鑑定工具缺乏。

特殊教育研究及各類特殊學生之鑑定需要有有效之研究及鑑定工具，但目前國內所使用的研究及鑑定工具絕大部分都相當老舊，而且一再重複使用。有的研究者使用國外所發展的工具來進行研究或鑑定工作時，不但未加以修訂，而且更是直接使用國外的常模。特殊學生之定義本來就不夠明確，再加上缺乏有效之研究及鑑定工具，因此，更難期望研究及鑑定結果能

符合科學研究最重要之條件：可重複(複製)性。未來在特殊教育方面的研究，除了需明確界定各類特殊學生之意義外，也要有系統的發展適用於國內的特殊教育研究及各類特殊學生之鑑定工具。

研究工具的品質不佳，除了會影響分類之一致性及精確性之外，也會影響所研究的變項(構念)彼此之關係大小的估計。信度越低，變項(構念)觀察分數彼此之關係就越小。為了提昇研究之品質及了解所研究的變項(構念)真正之關係的大小，除了努力發展高品質研究工具外，也可以使用功能較強的統計分析方法，來進行研究工具發展及研究結果之統計分析工作。如「結構公式模式」(structure equat- ion model，簡稱 SEM)可用來分析潛在構念彼此之關係及因果模式。另外，類推性理論(gen- eralizability theory)可分析來自於各不同面向(facets)的變異成分之大小，並據以設計適當的決策研究設計，那就可將測量誤差控制在可接受之範圍。

(四)研究對象難以符合大多數推論統計之基本要求：母群是常態分配的。

特殊教育研究的對象絕大部分屬於相當極端的特定群體，這些特殊群體的母群，其認知能力、情意或行為表現本來就較極端，不太可能符常態分配之要求。當然，推論統計其他要求，如變異數同質性、回歸係數同質性等也不太可能符合。除非所使用的推論統計分析方法之韌性很強，否則，所得結論可能不太可靠。

一般常用的推論統計分析方法，在母群嚴重違背上述假定時，各種參數之估計誤差的計算會有問題，因此，特殊教育研究需要有一些特殊的統計分析方法來估計各種參數之標準誤(standard error)，如：重複抽樣法(bootstrapping)。

(五)研究的變項之量尺水準未能符合等距量尺(interval scale)之要求。

此項限制不單是特殊教育研究獨有的問題，幾乎所有社會科學方面的研究所蒐集的變項(如智力、學業成就、數學態度等)之量尺水準，最多只能說介於次序量尺與等距量尺之間，並未達等距量尺之水準。大多數常用的統計方法(如平均數、標準差、t-test、ANOVA、相關及回歸分析等)均要求資料需符合等距量尺之要求，否則只能使用統計檢定效率較差的非母數統計檢定法(nonparametric tests)。

目前已漸普及的題目反應理論(item response theory，簡稱 IRT)，除了可以較有效解決測驗分數等化(test equat- ing)、電腦適性測驗(computerized adaptive testing，簡稱 CAT)、差別試題功能(differential item functioning，簡稱 DIF)等問題外，以 IRT 所發展的測驗，如果能符合所使用的 IRT 模式之假定，那所得到的題目參數及受試特質參數，可符合等距量尺之要求。

(六)適用於普通學生的能力測驗不一定也適用於特殊學生。

適用於普通學生的能力測驗，其題目難度通常較適用於能力接近平均值的學生，題目難度極高或極低者很少。特殊學生的能力範圍大多屬於能力量尺兩極端者，以適用於普通學生的能力測驗來測量特殊學生，常會有極高限效應(ceiling effect)或極低限效應(floor effect)之問題。也就是說因為試題難度與受試學生能力差距太大，以致於無法測出受試之真正能力水準。

為了能較有效測出特殊學生之真正能力水準，有必要發展適用於特殊學生的能力測驗。此種測驗最好以 IRT 來發展，題目的選取以能對特殊學生能力估計提供最大訊息者為依據。雖然特殊學生施測的題目及題數與一般學生不同，但因為每個題目均有 IRT 參數，所估計的能力仍然可和

一般學生比較，也就是說兩個群體所估計的能力是使用相同的能力量尺,這就是 IRT 在測驗分數等化之具體貢獻。

以 IRT 所發展的測驗，如果題庫夠大，題目難度分布範圍也夠廣，那還可以進一步發展成 CAT。CAT 有如爲每位受試分別量身訂製適合其能力水準之測驗，雖然每位受試施測的題目及題數均不相同，但還是可相互比較，而且測量所需時間較為經濟，測量精確性也可以達到事前預定之精確水準(如，能力參數估計之標準誤小於 0.3 個標準差，約相當於傳統信度係數 0.91)。

身心障礙特殊學生受限於某些方面之缺陷，其成長背景及感官經驗與一般學生常有很大之差異。各種能力測驗有很多題目對身心障礙的特殊學生可能不太公平，特別是對感官障礙的特殊學生更是如此。

前述之差別試題功能是指能力相近之幾個群體，在某些題目上的表現有明顯的差別。以往將此現象稱為「試題偏失」(item bias)，目前大部分學者以 DIF 來取代「試題偏失」。DIF 只是統計分析結果所得指數，如果進一步針對 DIF 指數較大的題目之內容詳細檢視後，發現那些試題確實包含與該測驗所要測量的構念或特質無關之成分，才能認定那些試題有偏失。目前較主要的測驗機構已將 DIF 檢定列為試題分析例行工作之一部分，針對身心障礙學生所發展的測驗，更應該進行試題之 DIF 檢驗。

(七)以傳統統計考驗來評估行為或心理治療之效果不一定有實質意義。

行為或心理治療效果的評估通常採用實驗的標準(experimental criterion)與治療的標準(therapeutic criterion)。實驗的標準是指比較實驗處理前與處理後效標行為(依變項)的差異情形。在大樣本的研究通常使用統計考驗方法比較實驗組與控制組平均數之差異，或是實驗組受試接受處理前與處理後平均數的差異，是否具有統計意義，如果有達顯著水準，實驗處理就被認為有效。治療的標準是指行為改變的方向與程度是否有價值與是否具有重要性，也就是實驗處理的效果在臨床上是否有意義，此即所謂的臨床意義或臨床顯著性(clinical significance)。更具體而言，臨床意義是指接受治療後，行為的改變是否讓受試回復常態(normality)，即回復到正常人的行為。一般而言，治療的標準似較實驗的標準難以達到，但就心理及行為異常治療的臨床研究而言，治療的標準可能比實驗的標準更重要。

以傳統統計考驗來評估實驗處理效果有兩個問題。第一，傳統統計考驗是在考驗實驗處理前後受試之平均得分的差異是否顯著的不同於 0，也就是平均而言，處理後是否確有變化，這對個別受試接受實驗處理的效果之個別差異沒有提供重要的訊息，而個別受試接受實驗處理之效果是否有所不同的訊息，對臨床學者而言是很重要的。第二，統計上實驗處理效果是否真正存在，對於臨床上是否有意義並沒有提供關鍵之資訊。統計上的顯著水準只是說處理的效果純粹是機遇造成的可能性很低，如小於.05，但是並沒有保證效果有多大，多重要，或是在臨床上是否有意義或足夠重要。行為或心理治療的功效(efficacy of psychotherapy)是指個案接受治療所能得到的實質助益，如，能使個案回歸正常生活。傳統統計考驗僅進行群體間之比較，對心理治療的效能所提供之訊息很少。

特殊教育實驗研究處理效果之評估，除了進行傳統統計考驗之外，也重視實驗處理效果量(effect size)之大小，以及實驗處理對受試是否有實質助益之評估。

## 二、單一受試研究資料分析方法之新趨勢

特殊教育研究屬於單一個案時間系列研究，其主要目的通常是在分析特殊學生在某些特質或行為方面之成長軌線，或是評估某些介入(實驗處理)是否有效。單一個案時間系列研究資料的性質不同於一般量化研究資料，一般常用的統計方法不太適用。目前使用最多的統計分析方法是C統計數。

### (一)評鑑「介入效果」的簡易時間系列分析(C統計數)

1.C統計數之意義及計算公式

vonNeumann,Kent,Bellinson,和Hart(1941)描述的兩種時間系列變異數估計值(二種互為正交，即沒有重疊)。公式1是第1種：

$$S^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2 / N \qquad \text{(公式 1)}$$

公式1是離均差平方的平均數。例如有3個資料：1，2，3，其平均數( $\overline{X}$ )為2，資料數(N)為3，因此

$$S^2 = [(1-2)^2 + (2-2)^2 + (3-2)^2]/3 = 0.67$$

如果有5個資料：1，2，3，4，5，其$\overline{X}=3$, $N=5$, $S^2=2$。

從上述兩個時間系列可看出：雖然「1，2，3」及「1，2，3，4，5」兩個時間系列的趨勢相同(均為直線趨勢且斜率均為1)，但觀察的時間點數不同，以離均差平方為依據所計算的變異數也會不同，這是因為觀察的時間點越多，平均的離均差平方越大。

公式2是時間系列相鄰值之差的均方(mean square successive difference，MSSD)，這是時間系列的第2種變異數估計值。

$$MSSD = \sum_{i=1}^{N-1}(X_i - X_{i+1})^2 /(N-1) \qquad \text{(公式 2)}$$

MSSD不受時間系列平均水準變動的影響。上述1，2，3和1，2，3，4，5兩個時間系列，其MSSD均為1，這是因為時間系列的趨勢相同，前後兩點之差距也會相同。

Tryon(1982)提出公式3的C統計數來評估某一時間系列是否有顯著改變。

$$C = 1 - \{[\sum_{i=1}^{N-1}(X_i - X_{i+1})^2]/[2\sum_{i=1}^{N}(X_i - \overline{X})^2]\} \qquad \text{(公式 3)}$$

公式 3 最右邊的分子與公式 2 類似，是時間系列相鄰值之差的平方和(沒有除以(N－1))，分母是離均差平方和之 2 倍，與公式 1 類似，但沒有除以 N，而是乘以 2。

公式 4 是 C 統計數的標準誤( $SE_c$ )之計算公式， $SE_c$ 只和資料數有關。

$$SE_c = \sqrt{(N-2)/(N^2-1)} \qquad \text{（公式 4）}$$

（公式 5 的 z 是 C 與 $SE_c$ 之比值）

$$z = C / SE_c \qquad \text{（公式 5）}$$

在資料數≧25 時，z 符合常態分配。資料數在 8-24 時 z 近似常態分配。即使是只有 8 個資料，也不會明顯偏離常態分配(Young,1941)。表 1 是樣本數在 8 至 25 時，達到.01 及.05 顯著水準的決斷值。表 2 是 C 及 z 之計算實例。

表 1　C 統計數之決斷值

| N | 顯著水準 | | N | 顯著水準 | |
|---|---|---|---|---|---|
| | .01 | .05* | | .01 | .05 |
| 8 | 2.17 | 1.64 | 18 | 2.25 | 1.64 |
| 9 | 2.18 | 1.64 | 19 | 2.26 | 1.64 |
| 10 | 2.20 | 1.64 | 20 | 2.26 | 1.64 |
| 11 | 2.21 | 1.64 | 21 | 2.26 | 1.64 |
| 12 | 2.22 | 1.64 | 22 | 2.26 | 1.64 |
| 13 | 2.22 | 1.64 | 23 | 2.27 | 1.64 |
| 14 | 2.23 | 1.64 | 24 | 2.27 | 1.64 |
| 15 | 2.24 | 1.64 | 25 | 2.27 | 1.64 |
| 16 | 2.24 | 1.64 | ∞ | 2.33 | 1.64 |
| 17 | 2.25 | 1.64 | | | |

*.05 顯著水準之決斷值均為 1.64

表 2　在 A－B－A 實驗設計使用 C 統計數的例子

| 階段 | 分數(X) | $D^2$ | C 統計數之計算 |
|---|---|---|---|
| | 28 | 324 | **A1 階段 C 之計算** |
| | 46 | 49 | $\Sigma D^2 = 1122$ |
| | 39 | 36 | SS(X)＝662 |
| 第一個 | 45 | 441 | $C = 1 - 1112/[2(662)] = 0.160$ |
| 基準線 | 24 | 16 | $SE_C = \sqrt{(10-2)/(10^2-1)} = 0.284$ |
| 階段(A1) | 20 | 225 | $z = 0.160/0.284 = 0.563$ |
| | 35 | 4 | |
| | 37 | 1 | |
| | 36 | 16 | |
| | 40 | 256 | |
| | 24 | 64 | |
| | 16 | 441 | **A1 至 B 兩階段 C 之計算** |
| | 37 | 64 | $\Sigma D^2 = 2762$ |
| | 45 | 729 | SS(X)＝4113.5 |
| | 18 | 1 | $C = 1 - 2762/[2(4113.5)] = 0.664$ |
| 團體 | 19 | 1 | $SE_C = \sqrt{(32-2)/(32^2-1)} = 0.171$ |
| 代幣 | 18 | 0 | $z = 0.664/0.171 = 3.883$ |
| 階段(B) | 18 | 25 | |
| | 13 | 1 | |
| | 12 | 9 | |
| | 15 | 4 | |
| | 13 | 4 | |
| | 15 | 1 | |
| | 16 | 25 | |
| | 11 | 9 | |
| | 14 | 0 | |
| | 14 | 4 | |
| | 12 | 1 | |
| | 13 | 1 | |
| | 14 | 9 | |
| | 17 | 1 | |
| | 16 | 1 | |
| | 15 | 36 | **B 階段最後 10 個加 A2 兩階段 C 之計算** |
| | 21 | 25 | $\Sigma D^2 = 353$ |
| | 16 | 49 | SS(X)＝441 |
| 第二個 | 23 | 9 | $C = 1 - 353/[2(441)] = 0.571$ |
| 基準線 | 20 | 36 | $SE_C = \sqrt{(20-2)/(20^2-1)} = 0.212$ |
| 階段(A2 ) | 26 | 0 | $z = 0.571/0.212 = 2.693$ |
| | 26 | 16 | |
| | 22 | 49 | |
| | 15 | 81 | |
| | 24 | | |

2.C 統計數的特徵

從公式 3 可看出當「相鄰的資料之差異值的平方和」是「離均差平方和」的兩倍時，C 公式右邊的分子與分母相同，C 就等於 0。此種狀況在時間系列資料相當靠近平均數時(沒有趨勢存在)最可能發生。當時間系列資料有任何趨勢或非靜態存在時，「離均差平方和」增加之速度要大於「相鄰的資料之差異值的平方和」，這使得 C 公式右邊的分數部份變小，而 C 變大。「離均差平方和」可反映所有類型的趨勢之存在，「相鄰的資料之差異值的平方和」則不受各類趨勢之影響。

此二者之比，類似變異數分析的 F 統計數。下列是兩個假想的時間系列資料：

1　2　3　4　5　6　(直線趨勢)
3　4　3　4　3　4　(無趨勢)

平均數均為 3.5，「離均差平方和」二者有很大差異，直線趨勢的有 17.5，無趨勢的只有 1.5。但就「相鄰值之差異值的平方和」而言，二者均為 5。直線趨勢的資料之 C＝1－5／(2(17.5))＝.857。無趨勢的資料之 C＝1－5／2(1.5)＝-.67。因此，無趨勢資料的 C 有可能是負的。另一個例子是在 1，2，3，1，2，3 之時間系列資料中。

$$C=1-8/[(2)(4)]=0$$

C 統計數是否達統計上的顯著水準，需視 C 是其標準誤($SE_C$)之多少倍(即 z)而定(參見公式 5)。$SE_C$ 完全是樣本數之函數，樣本越大，$SE_C$ 就會越小，例如，N＝8 時，Sc＝.31，N＝80 時，$SE_C$＝.11，N＝800 時，$SE_C$＝.04。只要樣本夠大，且 C 是正值，那所有的 C 之 z 考驗均會達顯著水準。統計考驗達顯著水準只是說效果是 0 的可能性很小，並不是說效果很大，這是所有統計考驗共同的限制。如果研究者希望知道效果是否在某一水準以上(如每天主動與人溝通的行為至少增加 3 次)，那可將實驗處理階段(如，表 2 的 B)的得分全部先減去 3，然後再進行 C 的 z 考驗，如果仍達顯著水準，那就表示溝通的行為至少增加 3 次。

另外，需值得注意的是，實驗效果是「正」或是「負」，C 統計數無法加以區分。C 是負的不是代表負效果。例如，在下列二個時間系列中，C 均為.56。

| | A 階段 | B 階段 |
|---|---|---|
| 正效果： | 5　6　5 | 8　9　8 |
| 負效果： | 5　6　5 | 2　3　2 |

實驗效果之正負需依據各階段的得分之平均數來判定。如果所有數據均相同，那公式 3 右邊的分數之分子及分母均為 0，無法計算 C 統計數。既然所有數據均相同，當然就沒有任何趨勢或效果存在，不用計算 C 統計數。

3.C 統計數之應用

C 統計數的目的是在回答所要分析的時間系列是否有任何趨勢存在，也就是說是否明顯不同於純隨機變異。C 統計數可用在下列幾方面：

(1)評鑑基準線階段(A1)的資料是否有趨勢存在，如無明顯趨勢再合併實驗處理階段資料進行趨勢考驗。基準線階段的 C 統計數最好沒有達統計上的顯著水準(如表 2 的 A1 )，這樣對實驗處理效果之評估較方便且精確。只要將實驗處理階段的資料加在基準線後面，一起計算 C 統計數並進行 z 考驗即可(如表 2 的 A1＋B)，如果達顯著水準就表示實驗處理的系列不同於基準線系列。

此種處理方式似乎不太合理，因為「A1」之資料數少於「A1+B」。資料數越少 $SE_C$ 就越大，即使趨勢一樣，「A1」要比「A1+B」難達顯著水準。解決此問題的方法之一，是採用相同的 $SE_C$（如 A1）。

(2)如果基準線階段就有趨勢存在，可採用下列方法評估實驗處理效果。

A.根據基準線之趨勢預測實驗處理階段之表現，如基準線為 1，2，3，4，5，具有直線趨向，斜率為 1，那可根據此直線趨向預測實驗處理階段的得分(如系列 1 的 B 階段)。

| | A 階段 | | | | | B 階段預測 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 系列 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | A 階段 | | | | | B 階段實際 | | | | |
| 系列 2 | 1 | 2 | 3 | 4 | 5 | 11 | 12 | 13 | 14 | 15 |
| | A 階段 | | | | | B 階段 | | | | |
| 系列 3 (2－1) | 0 | 0 | 0 | 0 | 0 | 5 | 5 | 5 | 5 | 5 |

B.將實際的時間系列(如系列 2)減去預測的時間系列(如系列 1)，然後計算差值系列(如系列 3)的 C 統計數。系列 3 為系列 2 減去系列 1。可看出 A 階段的起始效果為 0，B 階段表示實驗處理比自然發展趨勢要高出 6 分。接下來可計算系列 3 的 C 統計數並進行 z 考驗，結果 C＝0.8，z＝2.82，達.01 顯著水準。表示實驗處理階段優於純粹自然發展之趨勢。如果只用系列 3 的 B 階段(6，6，6，6，6)，或是只用系列 2 的「B－A」所得結果(10，10，10，10，10)來計算 C，那都無法顯示兩階段平均數之差異，只能顯示兩階段的斜率是否不同(即兩階段進步或退步的速率是否不同)。

(3)計算實驗處理階段最後 10 個資料之 C 統計數，以評鑑實驗處理是否已達穩定階段。如果實驗處理階段最 10 資料之 C 統計數未達顯著水準，可視為實驗處理效果已趨穩定(Tryon，1982)。接下來就可將實驗處理階段最後 10 資料與第二個基準線資料(如表 2 的 A2)合併，計算 C 統計，如達顯著水準(如表 2 右下的 C=0.571，z=2.693，P<.01)，就顯示實驗處理階段的表現與第 2 個基準線不同。

三、單一受試資料實驗效果量之計算

效果量(effect size)是在描述某樣本資料所呈現的效果之大小，它只代表兩組變項之關係強度。當樣本數相同時，效果量越大，代表自變項對依變項之影響程度越大。以往大多數研究者只重視統計考驗結果，近年來則逐漸轉而重視效果量。在 1994 年 APA 出版的出版手冊也鼓勵使用效果量來呈現研究結果。

以效果量來描述處理效果，不但具有統計考驗的多項優點，而且可避免統計考驗的多項問題。如能善用效果量，那就有下列優點。第 1 是在資料有趨勢或自我相關存在時，也可以計算效果量；第 2 是不同分析者可得到一致的結果；第 3 是著眼於處理與結果變項之關係強度，而非是否能拒絕虛無假設；第 4 是大多數效果量均容易計算，且易於解釋和了解，可用以補充其他資料分析法之不足。

效果量指數雖然與臨床意義(clinical significance)或社會效度(social validity)並非同義詞，但確實有某些關聯。如，本節的效果量指數d與Jacobson 和 Truax(1991)所提出的臨床意義指數RC很類似，RC=( $X_2 - X_1$ )/$S_{diff}$，$X_2$ 及 $X_1$ 分別為個案後測及前測的得分，$S_{diff}$是差異分數

之測量標準誤，$S_{diff} = \sqrt{2(S_E)^2}$，$S_E = S\sqrt{1 - r_{XX}}$，$S_E$是所使用的測量工具之測量標準誤，S是量尺的標準差，$r_{XX}$是信度。RC代表個案接受處理後，其得分的變化(即$X_2 - X_1$)是測量誤差($S_{diff}$)的幾倍，如果RC＞1.96 或RC＜－1.96，代表受試的得分確有變化，因為純粹是測量誤差造成的機率已小於.05。RC只是反映受試之得分是否確有變化，真正要具有臨床意義，應該是要回復一般人之行為表現水準。具體而言，受試之表現最好能比異常群體高二個標準差，而且不比普通人低二個標準差以上，其得分應較接近普通人之平均數。這些指數均有助於實驗處理效果是否具有實質意義之評估。

(一)效果量的計算公式

1.平均數改變量

d =() /$S_A$( $\overline{X}_A - \overline{X}_B$)　　(公式1)

$\overline{X}_B$及$\overline{X}_A$分別為處理期(B)及基準線期(A)依變項之平均數，$S_A$是基準線期依變項的標準差。例如表 1 之資料$\overline{X}_B$＝17.40，$\overline{X}_A$＝47.25，$S_A$＝3.77，因此，d=(17.40－47.25)/3.77=－7.92

表 1　基準線與處理期之基本統計

| 階段 | 平均數 | 標準差 |
|---|---|---|
| 基準線期(A) | 47.25 | 3.77 |
| 處理期　(B) | 17.40 | 7.00 |

d 適用於基準線與處理期均沒有明顯趨勢時，如果有明顯趨勢存在，那就不適用。

2.變異數改變量

有時候平均數雖然沒有明顯變化，但變異數卻有相當大的改變，此時就適合採用

「變異數改變量」的效果值。其計算公式分為兩個階段：

(1)計算代表兩個樣本變異數之比的F’ 統計數(大除以小)。

F＝$S_L^2$/ $S_S^2$　　(公式 2)

$S_L^2$及$S_S^2$分別代表兩個變異數中較大及較小者。

(2)計算代表效果值的$f^2$

$f^2 = (N_L - 1)$F’/( $N_L + N_S$ )（公式 3)

$N_L$及$N_S$分別代表較大變異數及較小變異數的資料點數。

表 2　基準線與處理期資料點數及變異數

| 階段 | 資料點數 | 變異數 |
|---|---|---|
| A | 9 | 345.36 |
| B | 6 | 44.81 |

表 2 中的$S_L^2$＝345.36，$S_S^2$＝44.81，故F’ ＝345.36/44.81＝7.71

另外，$N_L$=9，$N_S$=6，故

$f^2$=(9-1)(7.71)/(9+6)=4.11

3.有趨勢資料水準的改變量

基準線與處理期有明顯趨勢時，要評估兩個階段的水準改變量之大小，需用下列階層多元迴歸分析法。

(1)計算所有資料的一般最小平方迴歸方程式(兩個階段的水準當作沒有變化，合併計算)

Y=$b_0$+$b_1$T+e　　(公式 4)

Y是依變項值，T是觀察值的時間點順序(第 1 個資料之T為 0，第 2 個為 1，依此類推)。e是迴歸預測之殘差，$b_0$是依變項

之起始值，$b_1$是資料的趨勢之斜率(即迴歸係數)。

表 3　A B兩階段趨勢相同資料

| 階段(X) | 期間(S) | 時間點(T) | 反應(Y) |
|---|---|---|---|
| A | 1 | 0 | 2.620 |
| | 2 | 1 | 2.410 |
| | 3 | 2 | 2.980 |
| | 4 | 3 | 2.990 |
| | 5 | 4 | 1.250 |
| | 6 | 5 | 3.614 |
| | 7 | 6 | 2.586 |
| | 8 | 7 | 3.585 |
| | 9 | 8 | 3.383 |
| | 10 | 9 | 4.128 |
| B | 11 | 10 | 3.502 |
| | 12 | 11 | 2.960 |
| | 13 | 12 | 3.435 |
| | 14 | 13 | 3.225 |
| | 15 | 14 | 4.966 |
| | 16 | 15 | 2.630 |

Y＝2.52＋0.08T＋e

表 3 資料計算結果得$b_0$＝2.52，$b_1$=0.08，$b_0$=2.52 是圖 1 迴歸線的截距，也就是當T＝0 時(或S＝1 時)之Y的估計值，也就是迴歸線與Y軸之交點。b1＝0.08 ，是指T每增加 1 個單位，Y預期增加之量。依據該迴歸直線進行預測，其預測精確性指標($R^2$)為 0.2289。

(2)A 及 B 兩個階段分別計算個別迴歸線之斜率及截距。由於我們主要的興趣在描述平均水準(level)而非趨勢(trend)之變化，因此兩個階段迴歸線的斜率必須設為相等，只是截距不同而已。此種方程式可用下式表示：

$$Y=b_0^{'}+b_1^{'}T+b_2^{'}X+e \qquad \text{（公式 5）}$$

上述的X代表階段，以虛擬變項(dummy variable)的 0，1 編碼。如，0 代表處理A，1 代表處理B。其餘符號與公式(4)相同。圖 2 兩條迴歸線之$b_0$'=2.37，$b_1$'=0.13，$b_2$'=-0.54。其迴歸方程式為Y＝2.37＋0.13T－0.54X＋e。由於A階段之X＝0，故階段A之Y＝2.37＋0.13T＋e。階段B之X＝1，故階段B之Y＝2.37＋0.13T－0.54(1)＋e＝1.83＋0.13T＋e。此結果顯示B階段之截距比A小 0.54，也就是在調整「趨勢」之後，「水準」改變了－0.54(即 $b_2$'＝－0.54)。

(3)使用公式 5 和 4 的$R^2$之差距來計算與「水準」改變有關的效果量指數：

$$f^2=\ (R_5^2-R_4^2)/(1-R_5^2) \qquad \text{（公式 6）}$$

$R_5^2$和$R_4^2$分別是公式 5 及 4 之決定係數，在表 3 的實例中，公式 4 的$R_4^2$＝0.2289，公式 5 的$R_5^2$＝0.2612。調整趨勢後，與「水準」改變有關的效果量之$f^2$=(0.2612 － 0.2289)/(1 － 0.2612)=0.04。 此效果量指數代表效果很小。

(4)檢驗$R^2$之增加量或$b_2$是否達所指定之顯著水準。這是水準改變量之統計顯著性檢定，類似 C 統計數的考驗。

4.「斜率」(即趨勢)改變的效果量

基準線與處理期有明顯趨勢時，要評估兩個階段的趨勢改變量之大小，需用下列階層多元迴歸分析法。

階段間趨勢或斜率改變效果量的算法也是分為 3 個步驟，但是此處所指的效果量是指兩個階段個別迴歸線斜率與共同斜率之差異。

(1)計算和公式 5 相同的一般最小平方迴歸方程式之斜率及截距，也就是視同斜率沒有改變。

$$Y=b_0^{'}+b_1^{'}T+b_2^{'}X+e \qquad \text{（公式 5）}$$

依據表 4 的資料，得到

Y＝5.09－0.18T＋2.37X＋e

上式T的斜率($b_1'$)為－0.18，代表整體而言有隨時間而稍微下降之趨勢(參見圖 5)，也就是T每增加 1，Y就降低 0.18。X的斜率($b_2'$)＝2.37，代表B階段比A階段高了 2.37，這就是調整「趨勢」後，兩階段「水準」之差異(參見圖 3)。依公式 5 之迴歸方程式所得到的$R^2$是 0.4726。

(2)計算兩種迴歸線各有不同斜率之迴歸方程式。此方程式可用公式 7 表示：

$$Y=b_0''+b_1''T+b_2''X+b_3''XT+e \qquad \text{（公式 7）}$$

公式 7 增加了第 3 個加權項($b_3''$)，這樣就可以在單一直線方程式中呈現個別的斜率。依據公式 7 的圖 4 兩個階段的迴歸線斜率不同，圖 3 的斜率則相同，這是因為公式 7 增加了$b_3''$XT。B階段的X＝1，故$b_3''$XT＝$b_3''$T；A階段的X＝0，故$b_3''$XT＝0。A階段的迴歸線T的斜率為$b_1''$，B階段則為($b_1''+b_3''$)。以圖 6 之資料而言，Y＝4.73－0.08T+4.41X－0.25XT+e，故A階段斜率為(－0.08)，B階段為((－0.08)＋(－0.25)＝－0.33)，也就是接受B階段的處理後，行為頻率之下降趨勢較A階段快。依公式(7)得到的$R^2$＝0.5560。

(3)使用公式(5)及公式(7)所得到的$R^2$來計算與斜率改變有關的效果量。

$$f^2=\frac{(R_7^2-R_5^2)}{(1-R_7^2)}$$

從上述的$R_7^2$＝0.5560 及$R_5^2$＝0.4726，可得到

$f^2$=(0.5560－0.4726)/(1－0.5560)=0.19

$f^2$=0.19 代表效果量為「中等」，也就是足夠讓較小心的觀察者注意到效果之存在。這是指「趨勢」差異的效果量，而非「水準」差異的效果量。公式(7)代表「水準」改變的效果量部份是$b_2''$， 但因「趨勢」不同，「水準」之改變就比較沒有意義。

(4)檢驗$R^2$之增加量或$b_3$是否達所指定之顯著水準。這是趨勢改變量之統計顯著性檢定，類似 C 統計數的考驗。

(二)效果量的解釋

效果量要多大才具有實用上之意義實在很難決定。Cohen(1988, 1992)曾提出判斷效果量大小的大致原則給應用研究者參考。他所描述的「中效果量」(medium effect size)是指較謹慎的觀察者不必經過統計考驗，只憑肉眼觀察就可察覺效果存在，此效果量大致上相當於已出版的行為研究論文之典型效果量。「小效果量」(small effect size)比「中效果量」不易察覺效果之存在，但還不至於微不足道(trivial)。「大效果量」(large effect size)與「中效果量」之差距就如同「中效果量」與「小效果量」之差距一樣。如果可以假定母群體分佈之形狀(如常態分配)，那就可以用兩個群體之分佈重疊的比率來解釋效果量。如 d＝1，且兩個群體均為常態分佈時，其中高分組的平均數是在低分組平均數之上一個標準差之處，低分組有 84%低於高分組平均數。d 越大，兩個群體重疊的比率就越小。

在沒有「趨勢」時，用來說明「水準」改變的效果量指數d，其小、中、大效果量分別為 0.2，0.5，0.8(Cohen,1988, 1992)。以表 1 資料而言，d=－7.92，遠超過「大效果量」之標準。另一種效果量指數$f^2$，可在資料有趨勢存在時，用來描述趨勢及水準之改變，以及變異程度之改變。$f^2$為 0.02，0.15 及 0.35 分別代表小、中、大效果量(Cohen,1988,1992)。表 2

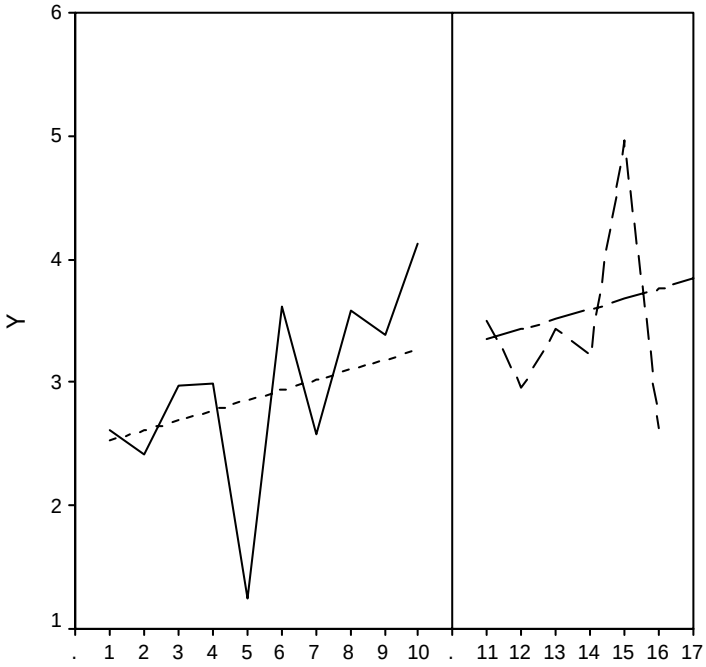


圖 1　表 3 資料兩階段迴歸線斜率及截距均相同

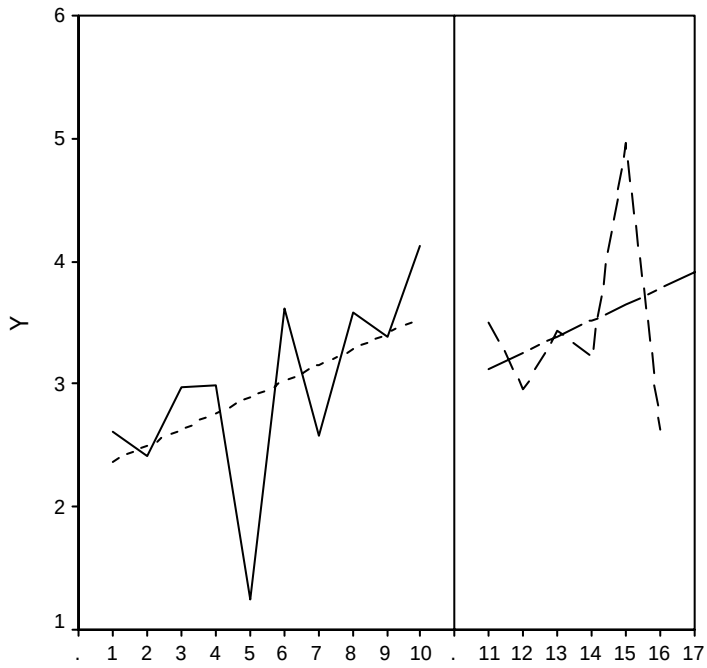


圖 2　表3資料兩階段迴歸線斜率相同截距不同

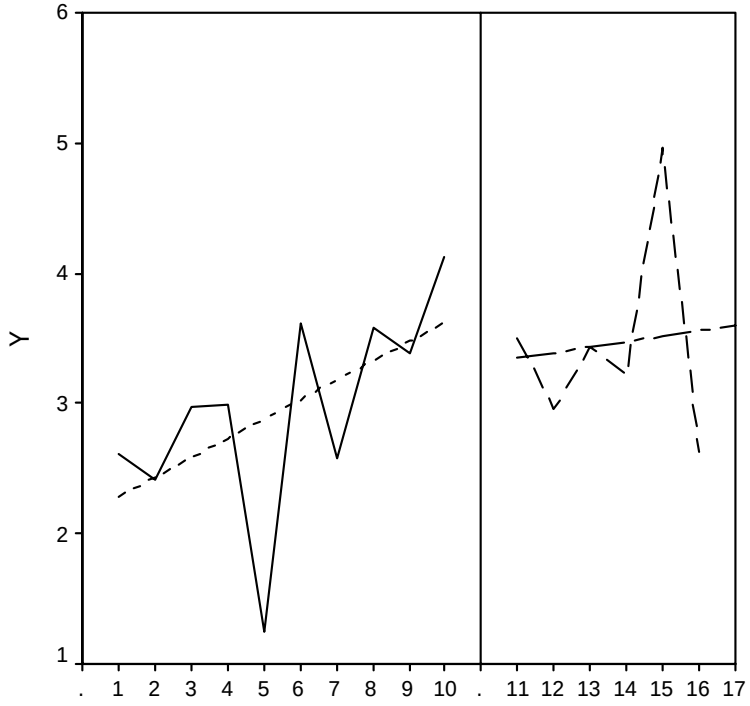


圖 3　表 3 資料兩階段迴歸線斜率及截距均不同

表 4　A 和 B 階段的趨勢有改變的資料

| 階段 | 階段虛擬變項(X) | 期間(S) | 時間(T) | X＊T | 反應(Y) |
|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 0 | 4.07 |
| | 0 | 2 | 1 | 0 | 5.60 |
| | 0 | 3 | 2 | 0 | 3.90 |
| | 0 | 4 | 3 | 0 | 4.90 |
| | 0 | 5 | 4 | 0 | 4.00 |
| | 0 | 6 | 5 | 0 | 5.00 |
| | 0 | 7 | 6 | 0 | 4.90 |
| | 0 | 8 | 7 | 0 | 3.30 |
| B | 1 | 9 | 8 | 8 | 6.70 |
| | 1 | 10 | 9 | 9 | 6.20 |
| | 1 | 11 | 10 | 10 | 5.90 |
| | 1 | 12 | 11 | 11 | 5.40 |
| | 1 | 13 | 12 | 12 | 4.80 |
| | 1 | 14 | 13 | 13 | 3.90 |
| | 1 | 15 | 14 | 14 | 5.50 |

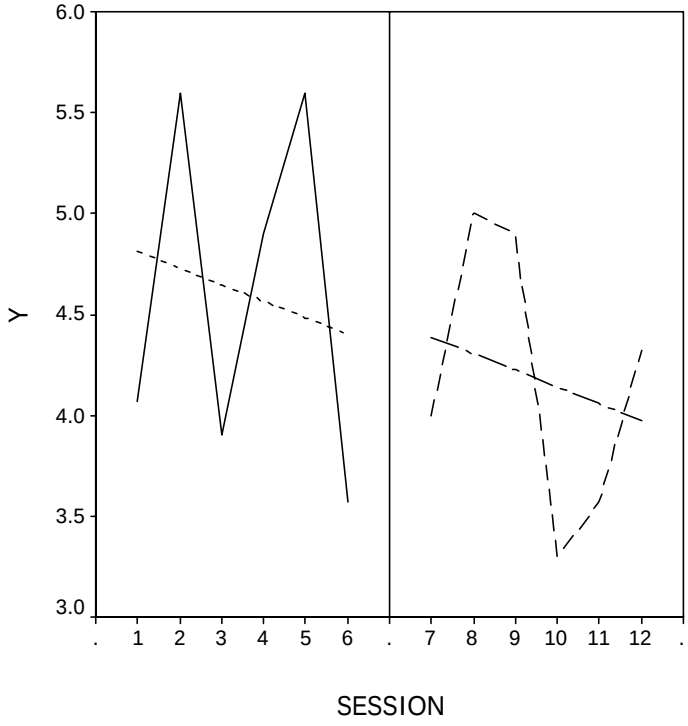


圖 4　表 4 資料兩階段迴歸線斜率相同截距不同

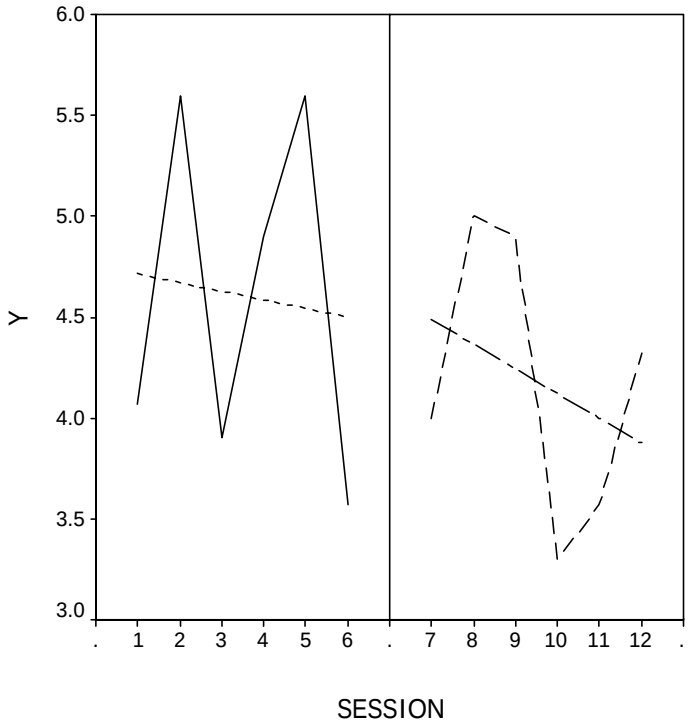

圖 5　表 4 資料兩階段迴歸線斜率及截距均不同

資料變異量改變的$f^2=4.11$，代表B階段的變異量和A階段比較有很大的改變。表 3 資料代表水準改變的效果量之$f^2=0.04$，屬於小效果量，表示水準沒有明顯之改變。另外，表 4 代表趨勢改變之效果量的$f^2=0.19$，屬於中效果量。$f^2$沒有指出關係的方向(正或負)，只告訴我們關係之強度，必須觀察樣本資料的變異量及迴歸線才能判斷改變之方向。d本身可以顯示改變之方向，d>0，代表B階段高於A階段，d<0 則代表B階段低於A階段。

## 結　論

以效果量指數來說明個案實驗研究之效果比圖示法更具體明確，但圖示法所傳達之訊息並非效果量指數所能取代。實驗效果之評估最好同時採用圖示法及效果量指數。效果量指數很容易計算，即使資料含有「趨勢」也可使用，不同研究者也可以得到相同的效果量指數，不像圖示法之解釋容易因人而異。效果量指數是用來說明實驗處理和結果變項間之關聯強度，而非在於拒絕虛無假設。單一受試實驗研究常用的簡易時間數列分析之 C 統計數，其著眼點是在拒絕虛無假設。本文雖然只舉例說明 AB 兩個階段的單一受試實驗效果量之計算方法，但是此方法也可應用在多個階段(如，ABA)之實驗情境。例如，在倒返設計(ABA)中，可分別計算 A→B 及 B→A 的效果量。在多基準線設計中，處理行為的改變，或研究對象的改變，均可計算每種行為或每個參與者的改變量。

效果量指數之應用與解釋需適切。當資料有趨勢存在時，如果使用 d 來描述水準之改變那就有問題，也就是會高估或低估效果量。例如，A 階段五次之觀察得到 1，2，3，4，5，B 階段 5 次之觀察得到 6，7，8，9，10。此資料有明顯之直線趨勢，隨時間從 1 遞升至 10，B 階段處理顯然對行為沒有什麼影響，只是依循原來 A 階段之趨勢直線上升而已。如果計算 AB 階段之 d 值，可得到

$$d=\frac{8-3}{1.4142}=3.5355$$

效果量相當大，如果下結論說B階段的處理造成行為很大的改變，那就大錯特錯了。如果改為比較兩個階段全體及個別迴歸線的截距及斜率，那就可以發現全體迴歸線與個別迴歸線完全重疊，顯示沒有任何改變。統計考驗之「顯著性」(significance)常被誤解，以為有達顯著水準(即拒絕虛無假設)代表效果很大。事實上，拒絕虛無假設只是說效果是 0 的概率很低，而非效果很大。另外，統計考驗受樣本數之影響很大，只要樣本夠大，任何考驗幾乎均會達顯著水準。即使常用的C統計數本身，也會受樣本數之影響。效果量統計數可以指出實際效果之大小，較具實用性及臨床應用價值。唯在應用效果量指數時需注意到樣本之大小，樣本太小很容易受誤差或機遇因素之影響，所得到的效果量就比較不可靠。筆者認為除了原有的統計考驗及效果量之計算外，可結合統計考驗之抽樣誤差及效果量兩種統計方法。如，先用抽樣誤差之概念計算出兩個階段平均數差異之 95%信賴區間，如$\bar{X}_B-\bar{X}_A$之 95%信賴區間為 5～10，$S_A=5$，那d之範圍為 1～2，也就是效果量在 1 至 2 之間屬於大效果量。

## 參考書目

Baer, D.(1977). Perhaps it would be better not to know everything. ***Journal of Applied Behavior Analysis, 10.***167-172.

Busk, P. L., & Serlin, R. (1992). Meta-analysis for single-case research: In T. R. Kratochwill and J. R. Levin (Eds.) ***Single-case research design and analysis: New directions for psychology and education*** (pp. 187-212). Hillsdale, NJ: Erlbaum.

Cohen. J.(1988). ***Statistical power anay- sis for the behavioral sciences*** (2nd ed.). Hillsdale . NJ: Eribaum.

Cohen. J.(1992). A power primer. ***Psy- cho- logical Bulletin, 112.***155- 159.

Crosbie, J. (1989). The inappropriate- ness of the C statistic for assess- ing stability or treatment effects with single-subject data. ***Behavi- oral Assess- ment,11,*** 315-325.

DeProspero, A., & Cohen, S. C. (1979). Inconsistent visual analysis of intrasubject data. ***Journal of App- lied Behavior Analysis, 12,*** 155-159.

Foster-Johnson. L. (1995). ***Methods of data analysis reported in 20 years of the Journal of Applied Behavior Anal-* ysis.** Unpublished manuscript. Univ- ersity of South Florida. Tampa.

Furlong, M. J., & Wampold, B. E.(1982). Intervention effects and relative variation as dimensions in exper- ts' use of visual inference. ***Jou- rnal of Applied Behavior Analysis,15,*** 415-421.

Gottman, J. M., & Glass, G. V. (1978). Analysis of interrupted timeser- ies experiments. In T. R. Kratoch- will (Ed.) ***Single-subject research: Strate- gies for evaluating change*** (pp. 197-235). New York: Academic Press.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statist- iccal approach to defining meanin- gful change in psychotherapy res- earch. ***Journal of Consulting and Cl- inical Psychology, 59 (1)***,12-19.

Jones, R. R., Weinrott, M. R., and Vaught, R. S. (1978). Effect of serial dependency on the agreement between visual and statistical inference. ***Journal of Applied Behav- ior Analysis, 11,*** 277-283.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in sin- gle-subject research. ***The Journal of Experimental Education,65(1)***, 73-93.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case ti- me series: Effects of variabili- ty, Seri al dependence, and magn- itude of intervention effects. ***Jo- urnal of Applied Behavior Analysis, 23,*** 341-351.

Michael, J.(1974). Statistical inference for individual organism research: Mixed blessing or curse? ***Journal of Applied Behavior Analysis, 7,*** 647-653.

Ottenbacher, K. J. (1992). Analysis of data in idiographic research: Issues and met- hods. ***American Journal of Physical Medicine and Rehabilitation, 71,*** 202-208.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.) ***Sin- gle-subject research: Strategies for evaluating change*** (pp. 101-162). New York: Academic Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data and current research into the stimuli controlling it. In T. R. Kratochwill and J. R. Levin (Eds.) ***Single-case research design and analysis: New directions for psychology and education*** (pp.15-40). Hillsdale, NJ: Erlbaum.

Skinner, B. F. (1963). Operant behavior. ***American Psychologist, 18,*** 503-515.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. ***Behavioral Assessment, 3,*** 93-103.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P.(1989). Applications of meta analysis in individual-subject research. ***Behavior Assessment, 11,*** 281-196.