Journal of Taiwan Normal University:
Education
2007, 52 (2) , 115-133

Meta-Analysis of the Pattern of Change in IQ Scores over a Three- Year Period for Exceptional Children with LD, EBD, and MR

Hsin-Yi Chen
Department of Special Education, National Taiwan Normal University

Jian-Jun Zhu
Frances Frey
Department of Psychometrics, Psychological Corporation, U. S. A.

Abstract

The techniques of meta-analysis were used to arrive at a quantitative synthesis of the results of 39 samples, based on the retesting of exceptional children using the WISC-R or WISC-III over an average time interval of three years. The primary findings indicated that IQ scores do change over time, and that the patterns of change for FSIQ, VIQ, and PIQ vary as a function of disability group. On average, children with LD and MR scored 2 points lower on retested VIQ, while retested PIQ increases significantly for LD and EBD children. Moreover, several factors including age, ability level, retest interval, and test version were found to be meaningful moderating factors which affect the variability of effect sizes for learning-disabled children. Further implications were also discussed.

Keywords: exceptional children, meta-analysis, retesting , WISC-III, WISC-R

Introduction

For diagnostic reasons, clinical individuals are frequently administered the same evaluations multiple times. Longitudinal and repeated diagnosis is also frequently conducted in the educational field, especially for educating exceptional children where individualization is the main concern. Thus, issues concerning retesting are of great practical importance (Horton, 1992; Kaufman & Lichtenberger, 2002).

Among various measures, Intelligence Quotient (IQ) has been one of the most frequently repeated measures in educational settings. Current research, which reports significant test-retest correlations, generally indicates stability of IQ over a period of three years in exceptional children (Canivez & Watkins, 1998; Sattler, 2001). However, these high correlations merely reveal similar rank orders between two sets of scores, they provide little information about ability level change. This is unfortunate, because change in ability levels is of great interest and practical importance to educational diagnosticians.

Findings on directions and magnitudes of IQ score changes across studies remain controversial. Canivez and Watkins (1998)

suggested that further investigation on existing discrepant results is needed. One attempt to investigate these discrepancies was a traditional literature review conducted by Sattler (2001) in which he summarized that the mean changes over a three-year period in Verbal, Performance and Full scale IQs are less than ±5 points in children with special needs. Despite the contribution of this literature review, we believe that a meta-analysis is needed to average results across studies in a quantitative manner (Hunter & Schmidt, 2004). In addition to investigating the mean ability level change, we were also interested in explaining the variability among score changes reported by different research. We hypothesized that it has systematic sources. For example, the pattern of IQ score change may be moderated by variables such as disability category, age (Horton, 1992; Mitrushina & Satz, 1991; Zhu, Woodell, & Kreiman, 1997), intellectual level (Bauman, 1991a, 1991b; Rapport, Brines, Axelrod, & Theisen, 1997), brain status and cognitive function (Kvale, 1987; Shatz, 1981), or length of retest interval (Schuerger & Witt, 1989).

Based on literature, there are certain factors that truly cause the test results to change from time to time. The real cognitive ability change is the major concern among all factors. As Kaufman (1994) once mentioned, three years is a long time, during which real change is likely to take place in cognitive abilities. Effects of some factors are random and unpredictable, such as measurement error or un-stability of mental status from time to time. Meanwhile, there are still other factors which are expected to causing score increase on retesting, such as practice effect (Anastasi, 1988), Flynn effect (Flynn, 1984, 1987; Kaufman, 1990), or regression toward the means (Bracken, 1988). In addition, the aforementioned factors may interact with each other.

To date, though this retesting issue is so frequently encountered in practical settings, very little empirical research has been published with respect to an integrative and synthetic examination of the pattern of IQ score change. The main purpose of the present study was to integrate discrepant findings on mean IQ level change in disabled children. This study also investigated possible moderator variables, an analysis which has been lacking in most investigations of the stability of Wechsler Intelligence scales.

Method

Meta-analysis is a statistical method for aggregating research findings across many studies that examine the same question (Hedges & Olkin, 1985). It is ideal for synthesizing research on this current issue. For this meta-analysis, an extensive search on IQs re-evaluation was conducted using PsycInfo database. For practical concern, we limited search to published or unpublished articles, but not dissertation or thesis. The reference sections of all obtained articles were also inspected. Re-evaluation results based on 39 samples of exceptional children (N=3,444), which were reported in studies from 1980 to 2003, were

collected and carefully reviewed (Appendix A listed the summary for these 39 independent samples). All children had been administered either the Wechsler Intelligence Scale for Children-Revised (WISC-R) or the Wechsler Intelligence Scale for Children-the 3rd edition (WISC-III) twice. The overall mean age for these children at the initial testing is 9 years and 10 months (ranges from 7 to 13 years).

In order not to confound later results, we set a prior to categorize all studies by disability category. A total of four categories were specified: Learning disabled (LD), Emotional/behavioral disabled (EBD), Mental retarded (MR), and unspecified combined clinical samples. Our main interest was in the first three categories. However, findings based on literature review revealed that some studies were done with samples in which children with different disability types were mixed. We decided to keep these combined samples for two reasons: (1) to make our literature review on disabled children as complete as possible; and (2) to provide researchers better ideas about how mixed samples could produce neutralized results.

Basic statistics such as Mean, SD and r were extracted from each of the 39 samples, missing correlation values were estimated using weighted average values across all samples in the same disability category. Effect sizes (ES) which measure unstandardized mean gain scores in Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scaled IQ (FSIQ) were computed (Lipsey & Wilson, 2001). Positive values of ES represent children scored higher on retest, whereas negative values indicate lower retest performances.

Instead of using a typical standardized ES index liked (Cohen, 1988; Hedges & Olkin, 1985), we chose to use an unstandardized ES, because the same Wechsler IQ scores are used in all selected research findings to be meta-analyzed. Findings from different samples are on the same Wechsler IQ scale (M=100, SD=15), and are therefore numerically comparable. Keeping the original IQ unit also makes interpretation of the results more intuitively meaningful.

To conquer the known variation problem due to sampling error, the inverse variance weight, which represents precision, was used to weight each ES. A weighted average of the effect sizes was computed for FSIQ, VIQ, and PIQ in each disability group. z statistic and 95% confidence intervals were analyzed to assess the direction and magnitude of the IQ score changes (Lipsey & Wilson, 2001: 113-115). Homogeneity analyses, which utilized Q statistics (Lipsey & Wilson, 2001: 115-116), which is distributed as a chi-square, were also conducted to determine whether groups of effect sizes were homogeneous. If a heterogeneous group was identified, the method of weighted multiple regression (Hedges & Olkin, 1985) was applied to identify possible moderator variables which could meaningfully explain excess variability.

Results

With an average test-retest interval of 2.88 years, mean Pearson product-moment correlation

coefficients, weighted by sample size (Hunter & Schmidt, 2004), for the stability of FSIQ, VIQ, and PIQ were: .80, .76, and .75 for the LD group; .88, .87, and .80 for children with EBD; .71, .68, and .69 for the MR group; and .84, .78, and .81 for unspecified combined clinical group. As expected, MR group show the lowest stability since IQ scores for this group varies the least. All coefficients were significant ($p<.01$), and confidence intervals presented in Table 1 indicated substantial long-term stability on relative ranking of IQs among exceptional children.

Table 1 Three-year test-retest stability of IQ scores: By disability category

		FSIQ	VIQ	PIQ
LD	$\overline{ES}$ [a]	.80**	.76**	.75**
	SE[b]	.03	.04	.04
	95% C.I.[c]	.74 — .86	.69 — .84	.67 — .82
EBD	$\overline{ES}$ [a]	.88**	.87**	.80**
	SE[b]	.02	.02	.03
	95% C.I.[c]	.84 — .92	.83 — .91	.74 — .86
MR	$\overline{ES}$ [a]	.71**	.68**	.69**
	SE[b]	.04	.05	.05
	95% C.I.[c]	.63 — .79	.58 — .77	.59 — .78
Unspecified	$\overline{ES}$ [a]	.84**	.78**	.81**
Combined	SE[b]	.03	.03	.03
Sample	95% C.I.[c]	.78 — .90	.72 — .84	.76 — .87

Note: Average retest interval is 2.88 years.

[a] Weighted mean effect size (weighted mean correlation coefficients)

[b] Standard error of the effect sizes

[c] 95% confidence interval

** $p<.01$ * $p<.05$

The meta-analysis results for IQ score change by disability category are summarized in Table 2. Findings revealed that the mean weighted effect sizes ($\overline{ES}$) vary by disability group.

IQ score changes and heterogeneity of variance

The group with LD, on the average, showed no meaningful change on FSIQ after 3 years ($\overline{ES}$ =-0.06). However, a significant negative value on VIQ ($\overline{ES}$ =-1.97, $p<.01$) and positive value on PIQ ($\overline{ES}$ =1.54, $p<.01$) were observed, in which neither 95% confidence intervals covered 0. This finding explains that regular LD children score 2 points lower on VIQ and 1.5 points higher on PIQ in retesting. The Q statistics revealed that all three of these effect size distributions are heterogeneous (Q=116.28, 130.58, 98.70, all $p<.01$), and therefore need to be further examined to identify possible moderating factors.

The EBD group demonstrated meaningful gains on both FSIQ ($\overline{ES}$ =1.72, $p<.01$) and PIQ ($\overline{ES}$ =3.24, $p<.01$), while no real change was found on VIQ ($\overline{ES}$ =0.50). Their FSIQ and PIQ

were, relatively, 1.7 and 3.2 points higher after 3 years. The null assumption of homogeneity was supported for all three distributions of IQ effect sizes (Q= 2.42 to 7.06), which revealed the coherence of effect sizes within each IQ construct for children with emotional/behavioral problems.

Findings for mentally retarded children yielded significant negative values on both FSIQ ($\overline{ES}$ =-0.70, p<.01) and VIQ ($\overline{ES}$ =-1.69, p<.01), and these two distributions were homogeneous (Q=11.04, 7.01, P>.05). Results showed that Full-scaled and Verbal intelligence for mentally retarded children degrade in 3 years. Meanwhile, the distribution of PIQ effect sizes was found to have mean value equivalent to 0 ($\overline{ES}$ =0.29), and significant variability across studies (Q=29.74, p<.01).

For the six combined samples, there seems to be no real mean change on FSIQ after 3 years ($\overline{ES}$ =0.13). However, a significant negative value on VIQ ($\overline{ES}$ =-2.09, p<.01) and a significant positive value on PIQ ($\overline{ES}$ =2.24, p<.01) were detected. Meanwhile, distributions of effect sizes were all found to be homogeneous in nature (Q= 0.68, 3.38, 6.80 respectively, all p>.05).

Table 2 Characteristics of effect size distribution: By disability category

Disability Category	No. of reports	N	Mean Age	Initial FSIQ		Dependent Variable		
						FSIQ	VIQ	PIQ
					$\overline{ES}$ [a]	-0.06	-1.97**	1.54**
LD	22	2441	9.49	90.59	SE[b]	(0.14)[a]	(0.16)	(0.19)
					95% C.I.[c]	-0.33 — 0 .21	-2.28 — -1.66	1.18 — 1.90
					Q[d]	116.28**	130.58**	98.70**
					$\overline{ES}$	1.72**	0.50	3.24**
EBD	5	163	9.66	94.02	SE	(0.51)	(0.56)	(0.68)
					95% C.I.	0.73 — 2.72	-0.60 — 1.60	1.91— 4.56
					Q	2.42	7.06	3.28
					$\overline{ES}$	-0.70**	-1.69**	0.29
MR	6	505	11.49	63.66	SE	(0.26)	(0.31)	(0.30)
					95% C.I.	-1.22 — -0.19	-2.28 — -1.09	-0.30 — 0.89
					Q	11.04	7.01	29.74**
Unspecified					$\overline{ES}$	0.13	-2.09**	2.24**
Combined	6	335	9.54	80.12	SE	(0.43)	(0.52)	(0.50)
Sample					95% C.I.	-0.71 — 0.97	-3.12 — -1.07	1.26— 3.22
					Q	0.68	3.38	6.80

Note: Average retest interval is 2.88 years.

[a] Weighted mean effect size (unstandardized mean IQ gain scores)

[b] Standard error of the effect sizes

[c] 95% confidence interval

[d] Q statistic which tests the hypothesis of homogeneity

** p<.01 *p<.05

Identifying Moderators of the Variance in IQ Score Change

In order to identify moderator variables which explain the heterogeneity among the effect sizes of the 22 LD samples, weighted multiple regression analyses were conducted on the LD data. Although the mental-retarded sample also showed significant variation on PIQ effect sizes, six effect sizes were not considered stable enough to produce reliable analysis on sources of variability. We believe the aforementioned mean effect sizes estimation for MR could be reliable since the overall numbers of MR children from these six samples were 505, which is a large enough size for central tendency estimate. We thus conducted regression analyses on the LD group only.

FSIQ

Table 3 summarizes the findings of the weighted regression analysis on FSIQ, where the dependent variable was the FSIQ effect size, and the four independent variables considered were: (1) age (in years) at the initial testing, (2) ability level represented by the FSIQ at the initial testing, (3) retest interval (in years), and (4) the selected test version (WISC-R vs. WISC-III). We suspected that there would be a non-linear relationship between each of the first three variables and the effect size, thus quadratic regression models were fitted. Table 3 presented the findings.

Table 3 Weighted multiple regression of **FSIQ** effect size for learning disabled children

Models	Variables	b	SE	z	R^2	SE_{est}	SS_R	SS_E
1					.631	1.58	68.55**	40.08**
	(Constant)	-32.77	8.16	-4.02**				
	Age	5.74	1.63	3.52**				
	Age*Age	-0.24	0.08	-3.00**				
2					.067	2.47	7.44*	103.37**
	(Constant)	-107.02	45.68	-2.34*				
	FSIQ	2.40	1.00	2.40*				
	FSIQ*FSIQ[a]	-0.01337	0.01	-1.34				
3					.214	2.25	24.83**	91.46**
	(Constant)	22.12	10.79	2.05*				
	Interval	-19.99	7.79	-2.57*				
	Interval*Interval	4.19	1.41	2.98**				
4					.161	2.27	18.77**	97.52**
	(Constant)	0.28	0.16	1.75				
	WISC-III	-1.35	0.31	-4.35**				

[a] at least 5 decimal points have to be kept in order to get correct estimation

As can be seen in Table 3, a total of four models were testified, each test one variable at a time. For each model, all the following statistical indices were presented: (1) unstandardized regression coefficient (b); (2) corrected standard error for each regression coefficient (SE); (3) significant test for each regression weight (z); (4) variance explained by model (R^2); (5) standard error of estimation of this model (SEest); (6) sum of squares due to regression (SS_R); and (7) sum of squares due to error or residual (SS_E).

Results in Table 3 suggested that each of the four variables explains significant amount of variations on FSIQ change scores in the LD population (all SS_R were statistically significant). The R^2 indices indicate that each model accounts for 6.7% to 63.1% of the total variation. Age, which alone accounts for 63.1% of the total variation, is the most crucial factor for explaining the variation in FSIQ effect sizes. Besides, the retest interval, which explains a meaningful 21.4% of the total variation, plays a salient role too. Our results also revealed an effect of test version on the magnitudes of FSIQ effect sizes. On the average, studies with WISC-III reported a mean effect size which is 1.35 points lower than the one being reported in studies with WISC-R.

Figure 1 illustrates the extent to which the magnitude of FSIQ score changes can fluctuate with age, FSIQ, and testing interval.

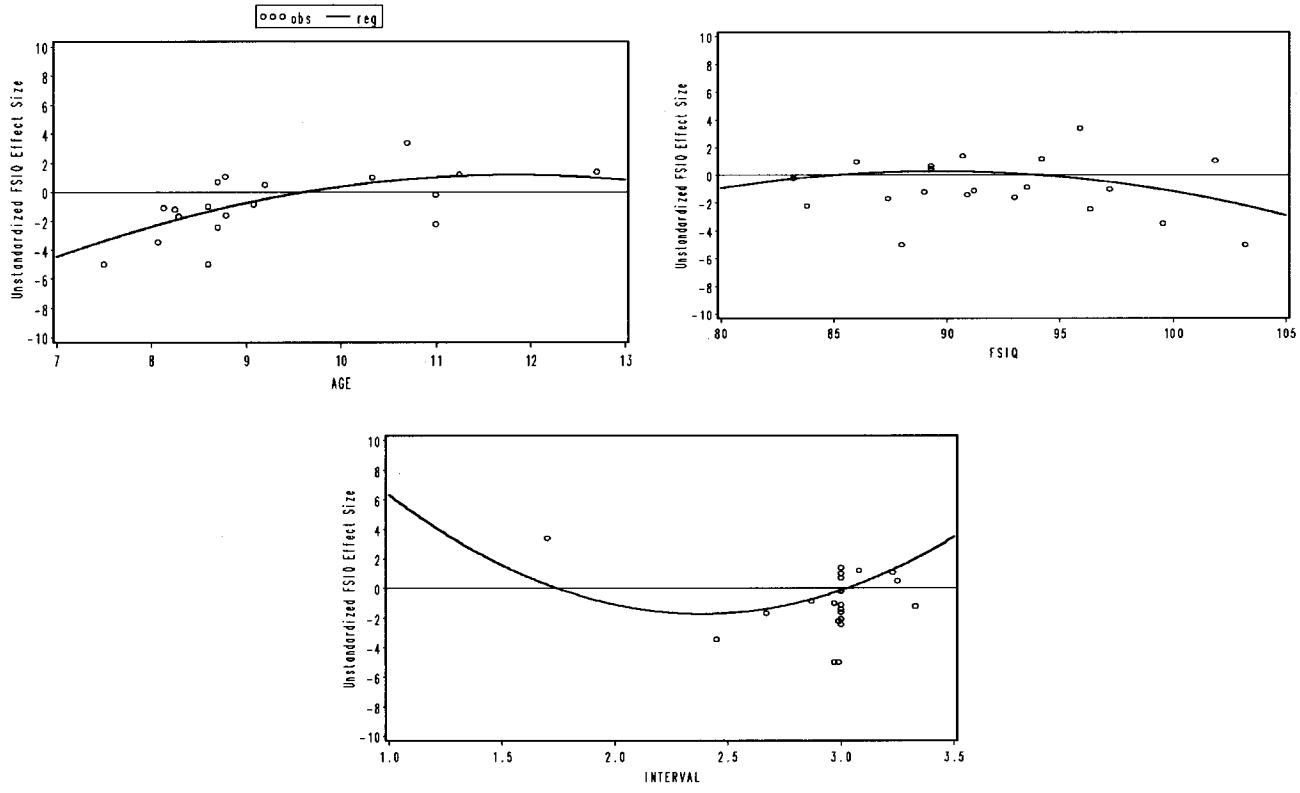

Figure 1 Effects of age, initial intellectual level, and retest interval on FSIQ effect size

Briefly, major findings based on the regression trends depicted in Figure 1 were: (1)

younger children tend to degrade more in retesting; (2) children with higher ability tend to show a more profound retest score decreasing; and (3) practice effect might be tremendous when retest within 1.5 years, while the interval effect diminished in 3 years.

VIQ

Regression analyses on VIQ effects sizes were summarized in Table 4. The same four independent variables were tested.

According to Table 4, age is the only significant variable which alone accounts for 42.8% of the total variation. The other 3 variables do not explain meaningful variations. Figure 2 performed this regression trend. Similar trend was scrutinized that VIQ of younger age LD children decrease more, while being retested in 3 years.

Table 4 Weighted multiple regression of **VIQ** effect size for learning disabled children

Models	Variables	b	SE	z	R^2	SE_{est}	SS_R	SS_E
1					.428	2.08	54.87**	73.21**
	(Constant)	-20.13	7.15	-2.82*				
	Age	2.90	1.46	1.98*				
	Age*Age	-0.10	0.07	1.42				
2					.015	2.65	1.87	126.24**
	(Constant)	-71.02	53.15	1.33				
	FSIQ	1.50	1.17	1.28				
	FSIQ*FSIQ[a]	-0.00815	0.01	-0.82				
3					.018	2.60	2.30	128.27**
	(Constant)	-17.07	14.29	-1.19				
	Interval	9.45	10.09	0.93				
	Interval*Interval	-1.47	1.79	-0.82				
4					.018	2.53	2.40	128.17**
	(Constant)	-2.12	0.19	-11.16**				
	WISC-III	0.56	0.36	1.56				

[a] at least 5 decimal points have to be kept in order to get correct estimation

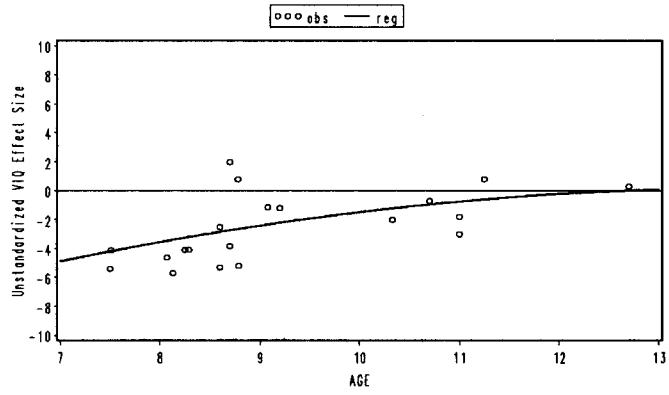

Figure 2 Effects of age on VIQ effect size

Results revealed that neither initial intellectual level nor testing interval were meaningful factors in explaining variability of VIQ effect sizes. The three-year reevaluation of VIQ seemed to yield a quite consistent mean effect size of -1.97, no matter how the initial FSIQ and testing interval varied.

PIQ

Table 5 summarizes the results of regression analysis on PIQ. All four variables were found able to explain significant amount of variability. Each variable accounts for 8.3% to 32.6% of the total variation.

Table 5 Weighted multiple regression of **PIQ** effect size for learning disabled children

Models	Variables	b	SE	z	R^2	SE_{est}	SS_R	SS_E
1					.128	2.20	12.14**	82.52**
	(Constant)	24.75	8.48	2.92**				
	Age	-5.00	1.73	-2.89**				
	Age*Age	0.26	0.09	2.89**				
2					.083	2.23	8.05*	89.39**
	(Constant)	-140.65	58.80	-2.39*				
	FSIQ	3.17	1.29	2.46*				
	FSIQ*FSIQ[a]	-0.01765	0.01	-1.77*				
3					.326	1.87	32.23**	66.48**
	(Constant)	52.61	10.35	5.08**				
	Interval	-40.38	7.66	-5.27				
	Interval*Interval	7.76	1.42	5.46**				
4					.304	1.85	30.04**	68.67**
	(Constant)	2.17	0.22	9.86**				
	WISC-III	-2.26	0.41	-5.51**				

[a] at least 5 decimal points have to be kept in order to get correct estimation

Comparatively, testing interval and testing version are two more salient moderators. Both explain more than 30% of the total variation. It was shown that WISC-R reports tend to generate significant positive PIQ effect (ES=2.17) for the LD population, while studies with WISC-III generate an estimated effect size value which is close to 0 (ES=-0.09). Interestingly, although age remains a significant moderator variable, its impact on PIQ is not as strong as the influences it makes on both VIQ and FSIQ effect sizes. Regression trends were shown in Figure 3.

According to Figure 3, main findings were: (1) both the effects of age and ability level on PIQ effect sizes are significant but trivial; (2) profound increasing in retesting was detected when retesting PIQ within 2 years.

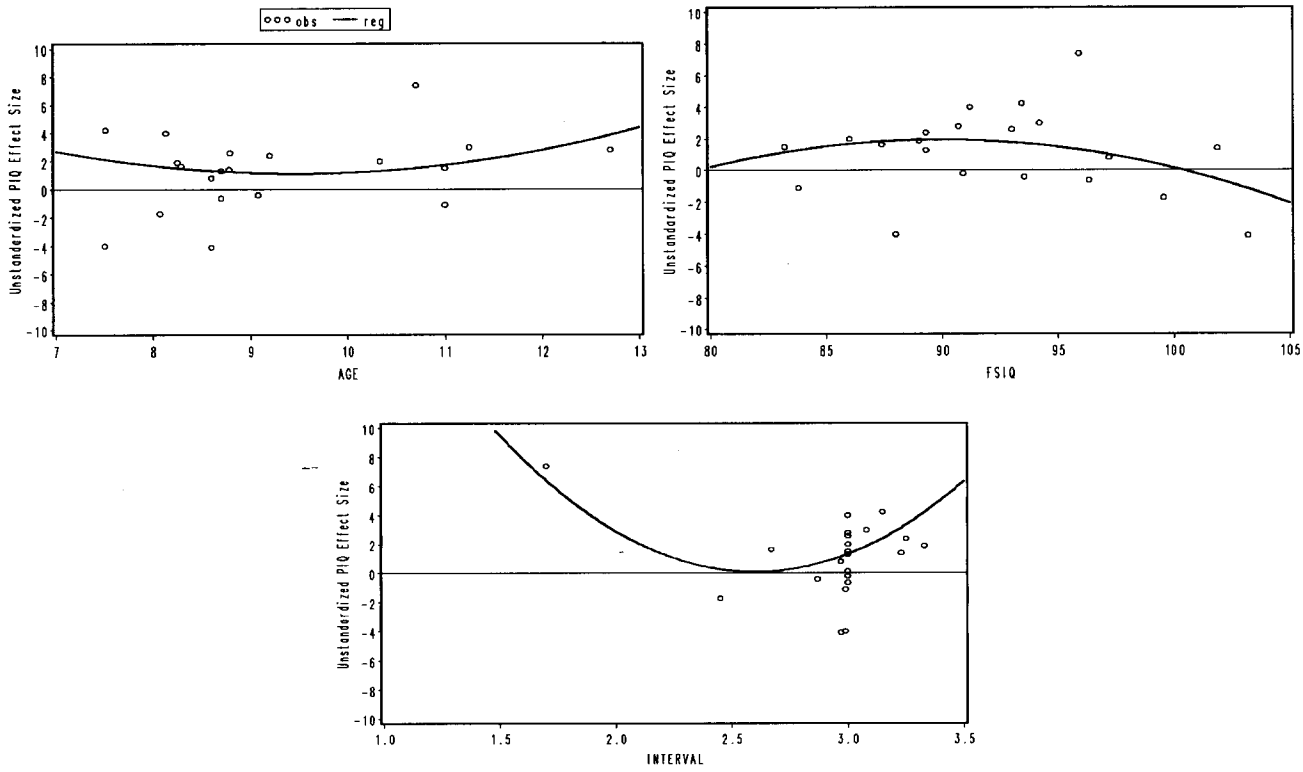

Figure 3 Effects of age, initial intellectual level, and retest interval on PIQ effect size

Discussion

The present study is the first to investigate systematically the mean and variability of effect sizes which represent IQ score change over three years in children with special needs. Results revealed that patterns of IQ change do vary by disability category. Mixed samples may provide neutralized results which make the differentiated effects remain in concealment, thus is not suggested for further applications.

Based on current findings, there was a tendency for the group with mental retardation to deteriorate over time with respect to both verbal and full scaled IQs. Learning disabled children also tend to show worse verbal ability at retest. However, they tend to earn significantly higher scores on performance scales which make the FSIQ remain unchanged. A different pattern was found with children who have emotional and behavioral problems. Their FSIQ and PIQ scores both gain significantly over 3 years, while verbal ability stays statistically the same.

Even though it is quite true that intelligence as a construct is presumed to be an enduring trait, practitioners should be reminded that IQ score is defined by a relative view-point. As a norm-referenced derived score, the magnitude of any person's IQ is scored by how much it deviates from the population mean. For disabled groups in this study, the change of their intellectual function over time cannot be

explained merely by natural progress, the relative standing of their IQ at different times could demonstrate real ability changes other than some known artificial effects. One note does deserve attention, other things being equal, effects such as practice effect, or regression toward the means all predicted a low ability children to show a higher score on retesting. Given the above general expectation, the decreasing trend revealed in this study really means some profound change. Since both MR and LD groups have known cognitive dysfunctions, which make it harder for them to learn. Relative deterioration in their cumulative learning is expected.

Our findings that both EBD and LD children gain significantly on PIQ at retest matches the truth that previous exposure dramatically alters children's performance on PIQ (Rapport et al., 1997) and practice effects are much greater for the performance scale than for the verbal scale (Kaufman, 1994). Practice effect reflects increased test sophistication, it is a combination of memory for item-specific and procedural aspects of the task, and may be detected even when parallel forms are used (Anastasi, 1988). It is said that instruments with speeded components are likely to result in significant practice effects (Dodrill & Troupin, 1975). Performance subtests are tests of adaptability and flexibility, their contents are mostly novel to children, and require children to utilize strategies to solve novel problems. Consequently, once a child has been tested, the novelty is gone. The reason for MR group not showing similar significant PIQ gain might be due to their known deficits in reasoning and learning (Harrison, 1990; Kamphaus, 2001; Mandes, Massimino, & Mantis, 1991). Shatz (1981) proposed that practice effects do interact with brain status, cognitively-impaired individuals will benefit less from practice than will normal individuals. Our results match this viewpoint.

While examining possible moderator variables which explain the variability of effect sizes in the LD population, we found that age is an important factor which shows significant impact on trends of FSIQ and VIQ effect sizes. Basically, the magnitude of decreases in FSIQ or VIQ scores was larger for younger subjects. This phenomena confirms Zhu, Woodell, and Kreiman's (1997) finding. Zhu and his colleague once explained this age effect in terms of differential developmental rates across age groups. According to them, the discrepancy between the slopes of developmental rate for normal versus LD children is larger at younger ages, and it gradually gets smaller at older ages. They used Figure 4 to demonstrate the hypothetical relationship between the development rate and age for an LD sample and a normal sample. In this figure, the difference between b and a simulates the score drop for young LD children who are tested 3 years apart, which is much larger than the distance between c and b, which represents the score drop for LD children who are first tested at age 9 and are retested when they are 12.

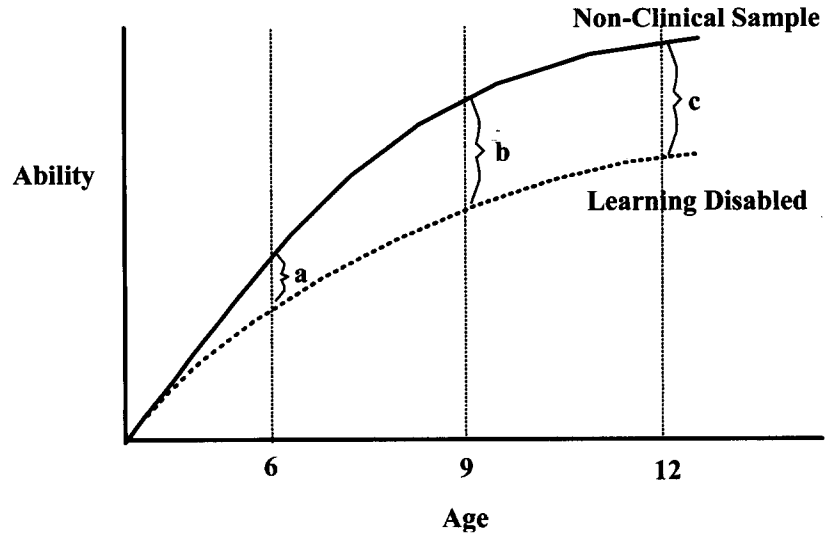

Figure 4 Differential Developmental Curves for Non-Clinical and Learning Disabled Children (Source: Zhu et al., 1997)

Besides, literature also suggests that practice effects could vary as a function of age (Horton, 1992). Ryan, Paolo, and Brungardt (1992) found that older adults benefit less than younger persons from retesting. It is highly possible that there exist age-related practice effects, and the age effects we detected in this study actually could be a combined result of both the aforementioned differential developmental slopes and the practice effects.

Another valuable finding was the differential effect which testing interval has on VIQ, PIQ, and FSIQ effect sizes. Basically, testing interval shows the strongest impact on PIQ effect size, while it has only small influences on FSIQ effect sizes, and plays no role in explaining the variability of VIQ effect sizes. These results again confirm the larger practice effect on PIQ than on VIQ. Practice effects on PIQ seem to keep functioning even after 3-years, whereas their effects on FSIQ seem to diminish after roughly 1.5 to 2 yeas for the LD population. Our preliminary finding does not quite support Canivez and Watkins (1998, 2001) suggestion that "practice effects seemingly disappeared when the retest interval was greater than 1 year." More data is needed for verification.

Finally, we found that among the 22 LD samples, the 15 ones which utilized WISC-R (n=1,777) showed mean effect sizes of 0.28, -2.12($p<.01$), and 2.17($p<.01$) for FSIQ, VIQ and PIQ respectively, whereas the values are -1.07($p<.01$), -1.56($p<.01$), and -0.09 for synthesizing the other seven WISC-III samples (n=664). The mean age (9.08 vs. 9.61 years), initial FSIQ level (92.17 vs. 90.05 IQ scores), and retest interval (3.03 vs. 2.89 years) between this two groups are comparable.

Interestingly, findings of these two versions (WISC-R and WISC-III) do indicate consistently regarding the direction and magnitude on VIQ mean effects size (-2.12 vs. -1.56, both $p<.01$), which shows that learning disabled children's performance do deteriorate over a period of three years with respect to their verbal abilities. On average, they get roughly 2 points lower on VIQ

in three years. However, there is a version effect on PIQ change score, where WISC-R studies reported an average of 2.17 points gained on PIQ, while WISC-III studies reported no significant gain. Similarly, there is also a version effect on FSIQ change score, where WISC-R studies reported no FSIQ change, while WISC-III studies reported an average of 1 point decrease on FSIQ over three years.

Both WISC-R and WISC-III are intelligence instruments with well-known good qualities (Kaufman, 1979, 1993, 1994; Sattler, 1992): besides the excellent norms which are population-representative (Wechsler, 1974, 1991), the internal reliability coefficients for FSIQ, VIQ, and PIQ are .96, .94, .90 in the WISC-R; and .96, .95, .91 in the WISC-III respectively. Contents of the WISC-R and WISC-III are quite similar, about 73% of the WISC-R items are retained in the WISC-III. The correlations between them are .89 for the FSIQ, .90 for the VIQ, and .81 for the PIQ.

There are several possible explanations toward current findings on version effect. The one we suspect is that the version effect is inextricability confounded with the recency effect. Generally speaking, WISC-III studies are conducted after the publishing year of 1991. The definition of "Learning Disability" itself has been recognized as a complex concept, people's knowledge and understanding about the nature of "Learning Disability" have improved over the past decades (Lyon, 1994), and are still making progress at the current time. Based on our hypothesis of recency effect, the cognitive nature of disabled "LD" samples collected by WISC-R may be different to some degree from the ones collected by more recent WISC-III reports. For example, the group of low achievers or slow learners might be better differentiated nowadays. Consequently, the samples in WISC-III reports might be much "purer" learning disabled children. They may have real learning problems, which make it harder for them to benefit from practice. This explanation matches our finding that WISC-III studies did not report a significant mean PIQ gain, while showing decreases on both FSIQ and VIQ scores. Nonetheless, this possible improvement in identifying "purer" LD merits continued investigation.

Another possible reason might be the growing slope for overall normal population is steeper than it was a few decades ago(Kaufman, 1994). This could also degrade the relative retest standing for LD children, while compared to the growing path of normal children.

Though our study did not find initial ability level a major factor in directing IQ scores change, we did find a similar trend to that described by Bauman (1991a, 1991b) such that there seem to be larger declines in children with above average IQs. We chose to be conservative in our interpretations of this current finding. More investigations are needed to further explore this issue.

Due to limitations of current study, interpretations and generalization of our results should be made with caution. First, our study integrated 39 samples of exceptional children. Among them, five reports with EBD (n=163), six with MR groups (n=505), and six with unspecified combined clinical samples (n=335) were included. Estimates on central tendency (like Mean effect size) based on these data

should be reliable, however, variation findings for these groups may be less stable, and especially more research on MR groups is needed to further investigate the variability of their PIQ effect sizes. Second, since majority of the available disabled samples are with retest intervals around three years due to practical constraints, the reported non-linear trend of interval effects should be treated only as a preliminary finding. Efforts to gain more reliable estimations on this issue should continue.

The findings in this study have important practical implications. Evidence based on this data indicates that patterns of IQ score changes vary by disability category. Previous research, which was based on mixed clinical samples (such as Canivez & Watkins, 1999; Elliott & Boeve, 1987; Vance, Hankins, & Brown, 1987) may get blended results with less meaningful interpretations. Our work revealed that upon retesting after three years, MR children tend to show downward FSIQ and VIQ trends, LD children also showed deteriorated performance on the VIQ. According to Flynn (1984), a one-point average increase in IQ score levels would be expected for normal children after the same retest period. This indicates that the gap between disabled children and normal kids could widen as they grow older. Practitioners are encouraged to consider these clinically meaningful group differences in addressing their differential developmental characteristics and needs. Besides, FSIQ is the average of VIQ and PIQ, if VIQ decrease and PIQ increase overtime, the FISQ will show no change. Therefore, merely examining FSIQ change is not quite meaningful. We recommended that more attention should be put on the VIQ and PIQ level.

In this study we also identified several factors, such as test version, initial testing age, and test-retest interval that moderate IQ score changes in learning disabled children. As Zhu et al. (1997) suggested, the score changes should be interpreted with extreme caution, because many factors and the interactions among them could contribute to observed differences. Further research is needed to validate these preliminary findings, and to continue investigating the nature and effect of moderating factors on IQ score changes.

References

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, P. L., Cronin, M. E., & Kazmierski, S. (1989). WISC-R stability and re-evaluation of learning-disabled students. *Journal of Clinical Psychology, 45*(6), 941-944.

Bauman, E. (1991a). Stability of WISC-R scores in children with learning difficulties. *Psychology in the Schools, 28,* 95-100.

Bauman, E. (1991b). Determinants of WISC-R subtest stability in children with learning difficulties. *Journal of Clinical Psychology, 47*(3), 430-435.

Bolen, L. M. (1998). WISC-III score changes for EMH students. *Psychology in the Schools, 35*(4), 327-332.

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26*, 155-166.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition. *Psychological Assessment, 10*(3), 285-291.

Canivez, G. L., & Watkins, M. W. (1999). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition among demographic subgroups: Gender, race/ethnicity, and age. *Journal of Psychoeducational Assessment, 17,* 300-313.

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition among students with disabilities. *School Psychology Review, 30*(2), 438-453.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dodrill, C. B., & Troupin, A. S. (1975). Effects of repeated administrations of a comprehensive neuropsychological battery among chronic epileptics. *The Journal of Nervous and Mental Disease, 161*, 185-190.

Elliott, S. N., & Boeve, K. (1987). Stability of WISC-R IQs: An investigation of ethnic differences over time. *Educational and Psychological Measurement, 47*(2), 461-465.

Finkelson, L., & Stavrou, E. (1999, April). *The stability of IQ in learning disabled students.* Paper presented at the Annual Convention of the National Association of School Psychologists, Las Vegas, NY.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171-191.

Haddad, F. A., Juliano, J. M., & Vaughan, D. (1994). Long-term stability of individual WISC-R IQs of learning disabled children. *Psychological Reports, 74*, 15-18.

Harrison, P. L. (1990). Mental retardation: Adaptive behavior assessment, and giftedness. In A. S. Kaufman (Ed.), *Assessing adolescent and adult intelligence*(pp. 533-585). Needham, MA: Allyn and Bacon.

Haynes, J. P., & Howard, R. C. (1986). Stability of WISC-R scores in a juvenile forensic sample. *Journal of Clinical Psychology, 42*, 534-537.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Horton, A. M. Jr. (1992). Neuropsychological practice effects x age: A brief note. *Perceptual and Motor Skills, 75*, 257-258.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed.). Needham Heights, MA: Allyn and Bacon.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York: John Wiley & Sons.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston, MA: Allyn & Bacon.

Kaufman, A. S. (1993). King WISC the Third assumes the throne. *Journal of School Psychology, 31*, 345-354.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: John Wiley & Sons.

Kaufman, A. S., & Lichtenberger, E. O. (2002). *Assessing adolescent and adult intelligence* (2nd ed.). Boston, MA: Allyn & Bacon.

Kaye, D. B., & Baron, M. B. (1987). Long-term stability of intelligence and achievement scores in special-learning-disabilities samples. *Journal of Psychoeducational Assessment, 3,* 257-266.

Kvale, V. I. (1987). WAIS-R practice effects. *Journal of Clinical and Experimental Neuropsychology, 9,* 35.

Lally, M. J., Lloyd, R. D., & Kulberg, J. M. (1987). Is intelligence stable in learning-disabled children? *Journal of Psychoeducational Assessment, 4,* 411-416.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: high for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology, 18,* 487-507.

Lyon, G. R. (1994). *Frames of reference for the assessment of learning disabilities*. Baltimore, ML: Paul H. Brookes.

Mandes, E., Massimino, C., & Mantis, C. (1991). A comparison of borderline and mild mental retardates assessed on the memory for designs and the WAIS-R. *Journal of Clinical Psychology, 47*(4), 562-567.

Mattison, R. E., Hooper, S. R., & Glassberg, L. A. (2002, December). Three-year course of learning disorders in special education students classified as behavioral disorder. *Journal of American Academy of Child and Adolescent Psychiatry, 41: 12,* 1454-1461.

Mitrushina, M., & Satz, P. (1991). Effects of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47,* 790-801.

Nichols, E. G., Inglis, J., Lawson, J. S., & Mackay, I. (1988). A cross-validation study of patterns of cognitive ability in children with learning difficulties, as described by factorially defined WISC-R verbal and performance IQs. *Journal of Learning Disabilities, 21*(8), 504-508.

Nydén, A., Billstedt, E., Hjelmquist, E., & Gillberg, C. (2001). Neurocognitive stability in Asperger syndrome, ADHD, and reading and writing disorder: a pilot study. *Developmental Medicine & Child Neurology, 43,* 165-171.

Oakman, S., & Wilson, B. (1988). Stability of WISC-R intelligence scores: Implications for 3-year reevaluations of learning disabled students. *Psychology in the Schools, 25,* 118-120.

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: the rich get richer. *The Clinical Neuropsychologist, 11*(4), 375-380.

Ryan, J. J., Paolo, A. M., & Brungardt, T. M. (1992). WAIS-R test-retest stability in normal persons 75 years and older. *The Clinical Neuropsychologist, 6,* 3-8.

Sattler, J. M. (1992). *Assessment of children* (Revised and updated 3rd ed.). San Diego, CA: Jerome Sattler.

Sattler, J. M. (2001). *Assessment of children: cognitive applications* (4th ed.). San Diego, CA: Jerome Sattler.

Schmidt, H. P., Kuryliw, A. J., Saklofske, D. H., & Yackulic, R. A. (1989). Stability of WISC-R scores for a sample pf learning disabled children. *Psychological Reports, 64,* 195-201.

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individual tested intelligence. *Journal of Clinical Psychology, 45*(2), 294-302.

Schwean, V. L., & Saklofske, D. H. (1998). WISC-III assessment of children attention deficit/hyperactivity disorder. In A. Prifitera & D. Saklofske (Eds.), *WISC-III clinical use and interpretation* (pp. 91-118). San Diego, CA: Acdemic Press.

Shatz, M. W. (1981). WAIS practice effects in clinical neuropsychology. *Journal of Clinical Neuropsychology, 3,* 171-179.

Smith, T., Smith, B. L., Barmlett, R. K., & Hicks, N. (1999, April). *WISC-III stability over a three-year period in students with learning disabilities.* Paper presented at the Annual Convention of the National Association of School Psychologists, Las Vegas, NV.

Stavrou, E. (1990). The long-term stability of WISC-R scores in mildly retarded and learning-disabled children. *Psychology in the Schools, 27,* 101-110.

Stavrou, E., & Flanagan, R. (1996, March). *The stability of WISC-III scores in learning disabled children.* Paper presented at the Annual Convention of the National

Association of School Psychologists, Atlanta, GA.

Truscott, S. D., Narrett, C. M., & Smith, S. E. (1994). WISC-R subtest reliability over time: Implications for practice and research. *Psychological Reports, 74*, 147-156.

Vance, B., Hankins, N., & Brown, W. (1987). A longitudinal study of the Wechsler Intelligence Scale for Children-Revised over a six-year period. *Psychology in the Schools, 24*, 229-233.

Vance, H. B., Blixt, S., & Ellis, R. (1981). Stability of the WISC-R for a sample of exceptional children. *Journal of Clinical Psychology, 37*(2), 397-399.

Webster, R. E. (1988). Statistical and individual temporal stability of the WISC-R for cognitively disabled adolescents. *Psychology in the Schools, 25*, 365-372.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised (WISC-R)*. San Antonio, TX: Psychological.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition (WISC-III)*. San Antonio, TX: Psychological.

Whorton, J. E. (1985). Test-retest Wechsler Intelligence Scale for Children-Revised scores for 310 educable mentally retarded and specific learning disabled students. *Psychological Reports, 56*, 857-858.

Wiese, M. J., Struer, J. H., Piersel, W. C., & Schwarting, F. G. (1987). Stability of the WISC-R and WRAT factor structures for an intellectually retarded sample: A 3-year follow-up. *Journal of Psychoeducational Assessment, 4*, 364-369.

Zhu, J., Woodell, N. M., & Kreiman, C. L. (1997, August). *Three-year re-evaluation stability of the WISC-III: A learning disabled sample*. Paper presented at the Annual Convention of the American Psychological Association, Chicago.

作者簡介

陳心怡，國立臺灣師範大學特殊教育學系，教授

Hsin-Yi Chen is a professor of Department of Special Education, National Taiwan Normal University, Taipei, Taiwan. E-mail: hsinyi@ntnu.edu.tw

朱建軍，美國心理公司心理計量部，主任

Jian-Jun Zhu is the Director of Department of Psychometrics, Psychological Corporation, U. S. A. E-mail: JJ_Zhu@harcourt.com

Frances Frey，美國心理公司心理計量部，研究助理

Frances Frey is a research assistant of Department of Psychometrics, Psychological Corporation, U. S. A. E-mail: Frances_Frey@harcourt.com

收稿日期：95.05.29
修正日期：95.09.29
接受日期：95.12.08

Appendix A Summary of the basic information for the 39 independent samples

Category	Sample Source	Sample size	Retest interval (year)
LD	Canivez & Watkins (2001)	406	2.87
	Nydén, Billstedt, Hjelmquist, & Gillberg (2001)	14	1.7
	Smith, Smith, Barmlett, & Hicks (1999)	54	3
	Finkelson & Stavrou (1999)	80	3
	Zhu, Woodell, & Kreiman (1997)-sample 1	33	2.99
	Zhu, Woodell, & Kreiman (1997)-sample 2	27	2.99
	Stavrou & Flanagan (1996)	50	3
	Truscott, Narrett, & Smith (1994)	107	3.25
	Haddad, Juliano, & Vaughan (1994)	402	3
	Bauman (1991a)	130	2.67
	Stavrou(1990)	100	3
	Schmidt, Kuryliw, Saklofske, & Yackulic (1989)	36	2.45
	Anderson, Cronin, & Kazmierski (1989)	113	3.33
	Webster (1988)	83	3.08
	Nichols, Inglis, Lawson, & Mackay (1988)	224	3.15
	Oakman & Wilson (1988)	150	3
	Lally, Lloyd, & Kulberg(1987)	60	3.23
	Kaye & Baron (1987)-sample 1	68	3
	Kaye & Baron (1987)-sample 2	31	3
	Whorton (1985)	221	3
	Mattison, Hooper, & Glassberg (2002)-sample 1	18	2.97
	Mattison, Hooper, & Glassberg (2002)-sample 2	34	2.97
EBD	Canivez & Watkins (2001)	47	2.81
	Nydén, Billstedt, Hjelmquist, & Gillberg (2001)	14	1.6
	Schwean & Saklofske (1998)	37	2.5
	Mattison, Hooper, & Glassberg (2002)	29	2.97
	Haynes & Howard (1986)	36	2.41
MR	Canivez & Watkins (2001)	66	2.9
	Bolen (1998)	70	3
	Stavrou (1990)	60	3
	Webster (1988)	72	3.08
	Wiese, Struer, Piersel, & Schwarting (1987)	148	3
	Whorton (1985)	89	3
Combined	Livingston, Jennings, Reynolds, & Gray (2003)	60	3.09
	Vance, Hankins, & Brown (1987)	32	3.1
	Elliott & Boeve (1987)-sample 1	56	3
	Elliott & Boeve (1987)-sample 2	56	3
	Elliott & Boeve (1987)-sample 3	56	3
	Vance, Blixt, & Ellis (1981)	75	2.17

師大學報：教育類
民國 96 年，52（2），115-133

學習障礙、情緒行為異常、與智能障礙學生間隔三年智力分數改變之統合分析研究

陳心怡
國立臺灣師範大學特殊教育學系

朱建軍　Frances Frey
美國心理公司心理計量部

摘　要

本研究目的在分析特殊兒童間隔三年後再測智力分數之變化模式，並探討相關影響因素。研究者根據 1980 年代後發表文獻內，共三十九組使用 WISC-R 或 WISC-III 為工具之特殊兒童智力再測樣本，以統合分析進行量化統整。主要研究發現如下：(1)間隔三年後，特殊學生智力表現的確有顯著變化；(2)智力再測分數之改變模式（方向與差異程度）依學生障礙類別而有不同，其中學障(LD)和智障(MR)學生的語文智商有下降趨勢，而學障與情緒行為異常(EBD)學生之作業智商則顯著上升；(3)學障兒童智力分數改變程度受到學生年齡、起始能力、再測間隔時間、以及測驗新舊版本等中介變項影響。文中並對未來研究方向提出建議與討論。

關鍵字：特殊兒童、統合分析、再測、魏氏兒童智力量表