Pergamon

Journal of School Psychology
41 (2003) 453–465

Journal of
School
Psychology

The rise and fall of IQ in special ed: Historical trends and their implications

Tomoe Kanaya[a,*], Stephen J. Ceci[a], Matthew H. Scullin[b]

[a] *Department of Human Development, Cornell University, MVR G 61, Ithaca, NY 14853, USA*
[b] *Department of Psychology, West Virginia University, Morgantown, WV 26506, USA*

Received 25 April 2003; received in revised form 6 August 2003; accepted 6 August 2003

Abstract

Using IQ records from 10 sites around the country, we examined longitudinal trends in IQ among mental retardation (MR) students during the Wechsler Intelligence Scale for Children (WISC) to WISC-Revised (R) transition in the 1970s and learning disability (LD) students during the WISC-R to WISC-3rd Revision (III) transition in the 1990s. Based on the work by Flynn [American Journal of Mental Deficiency 90 (1985) 236; Psychology, Public Policy, and Law 6 (2000) 191], we predicted a rise in IQ followed by a fall each time a newly normed IQ test is introduced into the schools and used to diagnose children as MR or LD. As expected, we found that mean IQs of MR and LD students followed a nearly identical parabolic trajectory, and differed depending on what year they were tested. IQs from the older norms were higher than IQs from the newly introduced norms. This systematic shift had a significant impact on the likelihood of a MR diagnosis. The broader implications of this pattern are discussed.

Keywords: Flynn effect; IQ; Mental retardation; Learning disability; Longitudinal studies

Introduction

The “Flynn effect” refers to the steady and substantial rise in IQ throughout the 20th century. It is now well established that when a standardized IQ test like the Wechsler Intelligence Scale for Children (WISC) or Stanford Binet is renormed, it becomes harder to obtain the same IQ on the renormed test (Flynn, 1984, 1987). For

* Corresponding author. Tel.: +1-607-255-7827; fax: +1-607-255-7856.
E-mail address: tk74@cornell.edu (T. Kanaya).

0022-4405/$ - see front matter
doi:10.1016/j.jsp.2003.08.003

example, an individual who receives a full-scale IQ of 100 on the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) would be expected, on average, to receive a full-scale IQ of 95 on the Wechsler Intelligence Scale for Children-3rd Revision (WISC-III; Wechsler, 1991) even when tested contemporaneously with the former (Flynn, 1998). Gradually, scores on the renormed version will rebound due to the steadily rising performance on IQ tests. However, initially the newly normed test will be associated with a substantial decline in IQs vis-à-vis its predecessor.

As a result, IQs have been following a rise-and-fall pattern each time a test is renormed. While the Flynn effect is well documented within the average range of the IQ distribution, less is known about its impact on those who score well below the mean, such as those in the mental retardation (MR) range, or individuals with specific cognitive deficits such as learning disability (LD). In this article, we explore the historical trends in IQ among MR and LD students during the WISC to WISC-R and WISC-R to WISC-III transitional periods.

The rise and fall of IQ over time: the Flynn effect and changing IQ norms

Although there are many intelligence tests available to school psychologists (e.g., Kaufman, Slosson), the Stanford-Binet and Wechsler scales are the most widely used IQ tests in the United States school systems today (Stinnet, Harvey, & Oehler-Stinnet, 1994). Using archived records, James Flynn, a political scientist from New Zealand, found "massive" IQ gains in America during the last century on these two tests (Flynn, 1984). More specifically, by systematically compiling and analyzing IQ data from over 7000 Americans who were tested on various versions of the Stanford-Binet and Wechsler tests, Flynn found a gain of approximately .311 points per year in the average IQ of Americans, adding up to over 3 points a decade. This systematic rise, dubbed the "Flynn effect" by Herrnstein and Murray (1994), has also been found in 20 industrialized nations around the world (Flynn, 1984, 1987, 1998).

This pronounced improvement in intelligence test performance over time causes IQ test norms to become obsolete within a generation or so. Intelligence test publishers are aware of this upward creep in IQ performance and respond by renorming their tests periodically. Whenever they renorm their tests, test-takers must attain higher raw scores or answer harder questions in order to obtain the same IQ as members of a prior cohort who took an older normed version of the test. Thus, over the course of the last century, the average IQ was periodically reset to 100 by making the IQ test substantially harder. Flynn found that whenever a group of subjects contemporaneously took both an older and a newly normed version of the Stanford-Binet or Wechsler series (e.g., the WISC and the WISC-R), it almost always scored lower on the newly normed test (Flynn, 1984). Therefore, Flynn not only discovered a rise in IQs over time, he also discovered a decline in IQs each time a test was renormed. Since the majority of individual IQ tests in the United States are administered to school children who are being considered for special education, Flynn (1985, 2000) has expressed concerns regarding the consequences of this systematic rise and fall in IQ on exceptional populations that rely so heavily on IQ, such as MR and LD.

Mental retardation

MR is characterized by significantly subaverage intellectual functioning (generally, an IQ of 70 or below is recommended), existing concurrently with limitations in conceptual, social, and practical adaptive skills (e.g., communication, social functioning, activities of daily living, etc.) by the American Association of Mental Retardation (2002). It is important to note that while the definition of MR has remained roughly the same over time, the social and political attitudes toward the MR have changed over the last century. In the early 20th century, many people believed the mentally retarded were dangerous or immoral, and feared they would dilute the overall genetic capabilities and morality of America (Gould, 1981). Indeed, in 1927, the Supreme Court upheld eugenics laws in the famous case of Buck v. Bell (1927).

After WWII, a movement towards improving the care of MR individuals began, resulting in large-scale deinstitutionalization and greater emphasis on the education of MR individuals. By the 1960s and 1970s, social policies and laws focused on the civil rights of groups who had been discriminated against in the past, including MR individuals (Heller, Holtzman, & Messick, 1982). This eventually led to Public Law 94-142 (1975), which guaranteed free, appropriate education for every handicapped child and a dramatic increase of people receiving MR services (Vitello & Soskin, 1985).

Throughout the 1980s, the number of students receiving MR services steeply declined. Explanations for this trend ranged from broader dissemination of research focusing on the negative impacts of labeling (e.g., Edgerton, 1967; Mercer, 1973), concern about the disproportionate amount of minority groups in the mild MR classroom (e.g., Burke, 1975; Mercer, 1973), and the advent of the LD diagnosis which siphoned off some individuals who might have previously been categorized as MR. Since the early 1990s, however, the number and percentage of students classified MR has leveled off and even increased slightly. Currently, there are approximately 600,000 MR students in the United States (US Department of Education, 2001). It is unclear whether this drastic reduction in MR diagnoses followed by a slight increase was driven by political issues (reluctance of psychologists to diagnose low scoring students as MR) or some other factor such as systematic fluctuations in IQ in accord with predictions from the Flynn effect. We shall return to this point below.

Learning disability

Unlike MR, LD has a relatively short history, and has experienced nothing but dramatic increases in "popularity" since its formal introduction by the National Advisory Committee on Handicapped Children (NACHC, 1968). While officially defined as "a disorder in one or more of the basic phonological processes involved in understanding or using spoken or written language" that are not due to visual, hearing, or motor handicaps, MR, emotional disturbance, or economic/social disadvantages (NACHC, 1968, p. 34), there is much controversy surrounding the LD diagnosis (e.g., Ceci, 1985; Singh & Beale, 1992). Part of this is due to variation in classification criteria between states and even among professionals within a state-although a popular (though controversial) operational definition is a significant discrepancy between an individual's IQ and achievement test score (Kavale & Forness, 2000). Regardless of which particular definition is used, an IQ in the normal range is an important component of the diagnosis. The LD diagnosis enables children to receive special instruction, extra time on tests, and other classroom benefits without having the stigma of an MR diagnosis. This may help explain the growing popularity of the LD diagnosis, which was initially estimated to affect 2% of the school-aged population (Public Law 94-142) but now affects nearly 6% of this population (U.S. Deparment of Education, 2001). Currently, there are nearly 2.9 million students receiving special education services for LD, making it the largest special education population (US Department of Education, 2001).

The implications of the Flynn effect on MR and LD students

Given how important IQ is to MR and LD diagnoses, the fact that IQ is rising and falling over time has important implications for students with, or being tested for, such diagnoses. Flynn's (1984, 1987) objective in his seminal literature reviews, however, was to discover historical IQ trends among Americans who score in the average range. In order to eliminate possible "eccentric" results that might distort his findings, he excluded scores from gifted and mentally retarded individuals (Flynn, 1987). In addition, while his sample covered a large range of years—between 1932 and 1978—his sample did not include LD individuals, as the LD diagnosis was not formally introduced into the school system until the late 1960s and did not become widely used until the mid 1970s (Kavale & Forness, 1992).

A few studies have noted a decline in IQ scores among children classified as MR once the WISC-III replaced the WISC-R. The general finding within MR samples has been replicated by other researchers, although the exact difference in full scale IQ scores between the WISC-R and the WISC-III varies between five and nine points (e.g., Bolen, Aichinger, Hall, & Webster, 1995; Slate & Saarnio, 1995; Vance, Maddux, Fuller, & Awadh, 1996). This suggests that the Flynn effect and changing IQ norms not only affect individuals in the average range of the IQ distribution, but also individuals with MR, and that the magnitude of the effect may be even larger among MR children than among children in the average range.

Even fewer studies have explored the potential role of the Flynn effect on LD students. A notable exception is Truscott and Frank (2001) who found similar decreases LD students tested on the WISC-R and WISC-III. They further concluded that LD diagnoses may be "substantially" impacted by the Flynn effect over the course of an IQ test's norming cycle.

While these studies are very important and useful, they all investigated the potential role of the Flynn effect on special education students through a relatively narrow lens. For example, these studies concentrated on a single geographic region. Given the variability of both MR and LD diagnoses among different school districts, it is important to look at several districts in order to make results more generalizable. In addition, previous researches on MR and LD samples have focused solely on a single transition (e.g., just the WISC to WISC-R, or just the WISC-R to WISC-III) or on a single population (e.g., just MR students). They did not investigate multiple transitions at the same time, thereby not utilizing a historical/longitudinal, systematic approach, which would seem to accord with the historical, systematic nature of the Flynn effect itself. Therefore, the historical trends in IQ among individuals with MR or LD were not charted by Flynn and are still unknown today.

The presence of the Flynn effect among MR and LD students implies that IQ is rising and falling among these individuals independently of real cognitive gains or losses. In practical terms, this would mean that an MR or LD diagnosis is dependent not only on the particular IQ an individual receives, but the year that individual is tested and the test norms used. Such a finding would call to question the use of IQ, and in particular, IQ cut-off scores, for such diagnoses.

Current study

The purpose of the present study was to explore longitudinal trends in the IQ of MR and LD students. Specifically, we examined students at their initial testing. At the outset of this study, we anticipated that the IQ of all children, including MR and LD students, would steadily rise until a new IQ test was introduced, and then fall upon the inception of the new norms. In other words, the IQ would be different depending on the year the student was tested and the test norms used: Students tested during the last years of an older test norm would have higher IQ than students tested during the first years of a newly normed test. If true, this could have a large impact on the actual diagnosis of both MR and LD, in opposite directions as will be seen. In the following analyses, we looked specifically at the WISC to WISC-R (published in 1974) transition, which occurred during 1972–1977, and the WISC-R to WISC-III (published in 1992) transition, which occurred during 1990–1995. Each of these transition periods represents the latest years of the older norm followed the earliest years of the newer norm.

Method

Sample

IQ data from 10,800 school psychologist special education assessments were collected from 10 different school districts across the country representing a diverse sample of geographical regions (Midwest, Southeast, West, South), neighborhood types (rural, urban, suburban), and subjects' socioeconomic status. Data included students' gender, age, testing date, IQ, test/testing norms used, and special education placement recommendations. Testing dates spanned approximately 43 years. Information regarding evaluation type (e.g., initial evaluation, retest, triennial reevaluation, etc.) was also obtained for most testings. Data were gathered by traveling to each school district and by manually recording all necessary information from each student's psychological testing file. This usually entailed two persons working full-time for at least 1 week to complete the transcriptions for a single school district.

In order to avoid any potential confounds between the effect special education services may have on students' IQ as well as potential pressure for retention in services despite scores that no longer qualified for such services, only testing records from initial evaluations and diagnoses (not re-evaluations) were considered in these analyses. Using this restricted criterion, our dataset only allowed us to analyze fewer than 800 students classified as MR during the WISC to WISC-R transition and fewer than 1000 students classified as LD during the WISC to WISC-III transition 20 years later. This is because there were relatively few instances of LD diagnoses in the 1970s and, likewise, relatively few instances of MR diagnoses in the 1990s (no doubt reflecting the historical contexts within MR and LD diagnoses mentioned earlier).

Ultimately, 768 MR students ($M_{IQ} = 64.17$, S.D. = 8.00) whose ages ranged from 6 to 16 and were tested during 1972 to 1977 from four of the school districts were extracted. Likewise, 923 LD students ($M_{IQ} = 92.71$, S.D. = 12.10) whose ages ranged from 6 to 16 and who were tested between 1990 and 1995 from nine of the school districts were also extracted and analyzed.

Results

Descriptive analyses

Table 1 summarizes the sample sizes and mean IQ for each year within the MR and LD samples. In both samples, the mean IQ dropped at the fourth year (1975 and 1993 for MR and LD students, respectively) and started to rise during the fifth and sixth years.

In both samples, a one-way Analysis of Variance (ANOVA) showed there was a significant difference between the years for both diagnoses (MR: $F(5, 768) = 6.04$, $p < .01$; LD: $F(5, 922) = 3.08$, $p < .01$. A Tukey's HSD test (displayed in Table 2) was performed on both groups to determine which specific years were significantly different from one another. As Table 2 illustrates, the only significant differences between years were in the fourth year of the transition periods. This fourth year represents the year after a newly normed IQ test was released. Among MR students the mean IQ in 1975 was significantly

Table 1
Descriptive statistics of MR and LD samples

Group	Year	N_s			TOTAL	M_{IQ}	S.D.
		WISC	WISC-R	WISC-III			
MR	1972	130	–	–	130	65.90	7.06
	1973	149	–	–	149	65.70	8.03
	1974	93	70	–	163	64.03	8.28
	1975	31	118	–	149	61.66	7.27
	1976	19	79	–	98	62.96	8.21
	1977	4	76	–	80	64.94	8.56
LD	1990	–	127	–	127	93.76	13.38
	1991	–	160	–	160	94.04	11.53
	1992	–	113	80	193	94.19	12.03
	1993	–	16	131	147	90.07	11.19
	1994	–	4	148	152	91.17	11.68
	1995	–	3	141	144	92.63	12.53

Table 2
Mean differences in IQ per year tested (in absolute values)

Group	Year tested	1	2	3	4	5	6
MR	(1) 1972	–					
	(2) 1973	.20	–				
	(3) 1974	1.87	1.67	–			
	(4) 1975	4.24*	4.04*	2.37	–		
	(5) 1976	2.94	2.74	1.07	1.30	–	
	(6) 1977	.96	.76	.91	3.28*	1.98	–
LD	(1) 1990	–					
	(2) 1991	.28	–				
	(3) 1992	.42	.14	–			
	(4) 1993	3.70	3.98*	4.12*	–		
	(5) 1994	2.59	2.87	3.02	1.10	–	
	(6) 1995	1.14	1.42	1.56	2.56	1.45	–

$^{*}p<.05$

lower than the mean IQs in 1972 ($p<.01$) and 1973 ($p<.01$), both of which represent the last 2 years of the WISC norming cycle. Among LD students, the mean IQ in 1993 was significantly lower than the mean IQ in 1991 ($p<.05$) and 1992 ($p<.05$). These years represent the last year of the WISC-R norming cycle and the first year of the WISC-III norming cycle, respectively. These results are in accord with our predictions and we will elaborate them below. But before doing so, it is necessary to clarify one aspect of the data that otherwise could cloud the findings.

As Table 1 illustrates, the mean IQ from 1992 was still heavily based on WISC-R test scores even though the successor norms (WISC-III) had been published. In addition, the old norms were not completely eliminated in either sample even after 3 years. In light of these observations, any significant differences are likely to be an underestimate of the renorming effect, given how gradual the Flynn effect rise is, the failure in many school districts to shift to newer norms when they first become available, and the ‘‘macro’’ nature of the above analysis. Next, we provide a finer-grained ‘‘micro’’ analysis that takes these factors into account. These analyses were run on Verbal and Performance IQ scores, in addition to Full IQ scores.

Group differences

In order to examine the impact of year tested on IQ on a ‘‘micro’’ level, we subdivided both samples, comparing the mean IQ for years 1972–1974 vs. 1975–1977 in the MR sample and the years 1990–1992 vs. 1993–1995 in the LD sample. The reasoning behind this was to compare the first half of the transitional time frame with the second half within both samples. The mean differences found were 2.3 Full Scale IQ points within the MR sample, and 2.75 Full Scale IQ points within the LD sample. *T*-tests revealed these differences were statistically significant (MR: $t(766)=3.97$, $p<.01$, d=.29; LD: $t(921)=3.47$, $p<.01$, d=.23.). In other words, the IQ of MR and LD students were significantly higher during the years immediately preceding a newly normed IQ test compared to the years immediately following a newly normed test, even though these clusters included many children who were tested with obsolete norms that work against the emergence of this finding.

This trend also held for Verbal and Performance IQs as well. Specifically, among the MR sample, the average difference in Verbal IQ was 2.41 points ($t(766) = 4.31$, $p < .01$, d=.32) and the average difference in Performance IQ was 2.64 points ($t(764) = 3.86$, $p < .01$, d=.28). Among the LD sample, the average difference in Verbal IQ was 2.31 points ($t(921) = 2.76$, $p < .05$, d=.18), while the average difference in Performance IQ was 2.87 points ($t(921) = 3.30$, $p < .05$, d=.22).

Finally, in order to fully explore the impact of changing norms on IQ over time, the MR and LD students were subdivided by the specific test norms that had been used for them. Therefore, the WISC scores were compared to the WISC-R scores within the MR sample, and the WISC-R scores were compared to the WISC-III scores within the LD sample. By doing this, we were able to eliminate any confounds created by the fact that different school districts replaced their old norms at different rates. Unsurprisingly, the results were very similar to the orthogonal contrasts described above. Specifically, the mean IQ on the old norms was significantly higher than the mean IQ on the new norms for both samples (MR; $t(766) = 5.42$, $p < .01$, d=.39, LD: $t(920) = 3.80$, $p < .01$, d=.25). The average difference between IQ was approximately three points for both samples. See Tables 3 and 4 for further details regarding these analyses.

Again, similar patterns were revealed in the Verbal and Performance IQs. Among the MR children, the average WISC to WISC-R difference Verbal IQ was 3.18 points ($t(766) = 5.79$, $p < .01$, d=.42), and 3.40 points in Performance IQ ($t(764) = 5.03$, $p < .01$, $d = 36$). Among LD children, the average WISC-R to WISC-III difference Verbal IQ was

Table 3
Mean differences (D) of IQs among MR students by years tested and test norms used

Years tested/norms used	N	M_{IQ}	S.D.	D
1972–1974				
Verbal IQ	441	68.27	7.80	
Performance IQ	441	67.96	9.24	
Full IQ	441	65.15	7.89	
1975–1977				
Verbal IQ	327	65.86	7.46	– 2.41**
Performance IQ	325	65.32	9.52	– 2.64**
Full IQ	327	62.85	7.97	– 2.30**
WISC				
Verbal IQ	420	68.68	7.34	
Performance IQ	420	68.37	9.03	
Full IQ	420	65.57	7.20	
WISC-R				
Verbal IQ	348	65.50	7.87	– 3.18**
Performance IQ	346	64.98	9.62	– 3.40**
Full IQ	348	62.48	8.60	– 3.09**

** $p < .01$

Table 4
Mean differences (D) of IQs among LD students by years tested and test norms used

Years tested/norms used	N	M_{IQ}	S.D.	D
1990–1992				
Verbal IQ	480	93.89	12.87	
Performance IQ	480	95.43	13.29	
Full IQ	480	94.03	12.22	
1993–1995				
Verbal IQ	443	91.57	12.53	− 2.31*
Performance IQ	443	92.56	13.06	− 2.87*
Full IQ	443	91.28	11.83	− 2.75**
WISC-R				
Verbal IQ	412	93.91	13.21	
Performance IQ	412	96.11	13.63	
Full IQ	412	94.39	12.64	
WISC-III				
Verbal IQ	510	91.88	12.31	− 2.03*
Performance IQ	510	92.40	12.72	− 3.71**
Full IQ	510	91.36	11.50	− 3.03**

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

2.03 points, ($t(920) = 2.41$, $p < .05$, $d = .16$), and 3.71 points in Performance IQ ($t(920) = 4.27$, $p < .01$, $d = .28$).

The impact of IQ shifts over time

The group shifts in IQ were further explored in order to determine the impact of such changes on special education students. The difference of 2.3 IQ points (from the contrasts) and 3.0 IQ points (between the WISC and WISC-R) equated to 29% and 39% of the standard deviation within our MR sample, respectively. Of interest was the effect of adding/subtracting 3 IQ points to scores; for instance adding 3 points to an IQ of 68 to become a 71, or adding 3 points to an IQ of 69 to become a 72. When *all* 217 students who were initially tested between 1972 and 1977 with IQ between 68 and 73 (exactly 3 points on both sides of the 70 IQ cut-off) were isolated, 154 of the students with IQs between 68 and 70 (71%) were diagnosed MR. On the other hand, only 54% (98 of the 180 students) with IQs between 71 and 73 were diagnosed MR. Chi-square analyses confirmed that these proportions were statistically different ($\chi^2 = 11.59$, $p < .01$). Therefore, a three-point shift can greatly affect the probability of an MR diagnosis, particularly for those on the cusp of the cut-off. Also, as we have argued, this three-point shift is largely independent on any real cognitive change.

Likewise, the 2.75 and 3.0 IQ points found within the LD sample represent 23% and 25% of the LD sample's standard deviation, respectively. Unfortunately, given that there is no cutoff IQ for LD, but rather the diagnosis is often determined by a discrepancy between one's IQ and achievement score (which often varies from district to district), it is difficult to accurately quantify the impact of these IQ shifts among LD students. In locations where diagnostic criteria for LD can be held constant, fewer children would qualify for LD services because it would become harder to obtain a significant discrepancy between IQ and achievement scores. Clearly, additional research on this issue is needed.

Discussion

Flynn (1984) found the IQs of average Americans rose and fell over time, in a yo-yo like fashion due to systematic rises in IQs as test norms aged and subsequent declines in IQs following the renorming of IQ tests. Using a geographically and economically diverse sample of students tested for special education, the present study explored historical trends in IQ among MR and LD students tested on the Wechsler series of intelligence tests, the most widely used tests in American schools. Results revealed a similar trend to that discovered by Flynn for the middle part of the normal IQ distribution, but this time for students closer to the tails of the distribution. Specifically, a parabolic pattern was found in IQ among students initially diagnosed as MR and LD. In other words, the average IQs of MR and LD students depended greatly on the year tested and test norms used. The impact of this IQ shift—in the absence of any real shift in the students' actual cognitive ability—resulted in large shifts in MR classifications.

Analyses revealed that IQs were higher during the last years of an old norm compared to the first years on a newly normed test by a little less than three IQ points for both groups. In addition, more specific analyses revealed the average difference in IQ between students tested on the old versus norms was a little over three points for both groups. The slight disparity between these two findings was due to differing rates of overturn from old norms to new norms within and between school districts. The similarities in results between the two samples were fairly remarkable given they are almost 20 years apart, are based on different parts of the IQ distribution, and represent very different cognitive populations.

In the past, several studies have reported a decrease in IQ when a newly normed test was administered with students with cognitive deficits using test–retest data on the same students (e.g., Slate & Saarnio, 1995; Spitz, 1983; Vance et al., 1996). Only a few studies (e.g., Gaskill & Brantley, 1996; Spitz, 1989; Truscott & Frank, 2001) have linked their results to the Flynn effect. The present paper differs from these previous studies as well as from Flynn's analyses. Rather than relying on test–retest performances from the same students, the trends of IQ among *newly diagnosed* students were investigated and compared over time. While this is not an unequivocally superior approach to the one that was taken previously, this approach explored this trend through a year-by-year lens (which is important given how systematic the Flynn effect is). In addition, by using newly diagnosed students, any potential confounds that may arise from the impact special education services may have on an individual's IQ, as well as individual variations that may occur within this impact were avoided.

As previously mentioned, Flynn (1984, 1987) did not include the MR or the LD in his large, systematic analyses of IQ gains over time. All the published research done on MR and LD individuals has used comparatively small sample sizes and was unable to disentangle IQ changes from curricular changes. These data represent a wide array of geographic locations, neighborhood compositions, and socioeconomic status, making our results more generalizable than those gathered from previous research. In addition, previous research on MR and LD samples has focused solely on a single transition (e.g., just the WISC to WISC-R, or just the WISC-R to WISC-III). They did not investigate multiple transitions at the same time. Therefore, the repetitive nature of this trend—and thereby, the implication that it can be expected again—was lost.

Against these strengths, however, we frequently only had access to the psychologists' recommendations for each student's placement. Although in most cases, psychologists' recommendations are congruent with the final placement recommendations for a student, at times the placement committee may override the psychologist's recommendation or a parent may place pressure on the committee so that their child does or does not receive services. Such instances, however, occur very rarely, do not bias the results in a given direction, and would not change the overall trends found within our analyses. In addition, although it is not possible to be certain that these IQ trends in MR and LD recommendations are due solely to the Flynn effect and changing IQ norms as opposed to changing special education policies due to amendments in IDEA, funding fluctuations, or other cohort effects, the patterns that were found here are exactly what one would predict assuming a yo-yo trend in IQ anticipated on the basis of the Flynn effect. It is also important to note, that while the overall sample sizes for the group differences were large (over 300 in each group), the sample sizes for each individual year were comparatively smaller (e.g., $n = 80$).

As we are coming to a close on the WISC-III norming cycle (the WISC-IV was standardized using a 2002 sample and was released the summer of 2003), the main conclusion that can be drawn is that the parabolic trend in IQ observed in these analyses can be expected again immediately in all individuals, including students being tested for special education services. In other words, we shall once again witness large numbers of American students' diagnoses changing without any true change in their cognitive ability (i.e., a student whose WISC-R IQ was 72 in 1992 and a peer whose WISC-III IQ was 69 a year later are each answering equivalent numbers or equivalent difficulty levels of IQ questions correctly, hence, no cognitive differences despite the three-point difference in their IQ and the potential for dramatically differently diagnoses).

The fact that the old IQ norms were not completely eliminated even after 3 years of the introduction of the new norms is also an important issue. This finding is not surprising, given that not all psychologists and school districts replace their IQ tests immediately after a newly normed test is released. IQ tests are expensive, and school budgets are often not able to provide new test kits to all of their testers at the same time. Also, some psychologists and districts may prefer not to use a newly normed test until all of the older scoring booklets are used up, so it may take many years before an older IQ test is completely phased out of a school system. In our experience, before an old test is completely phased out, different children may be tested on different norms in the same year—even within the same school district. Therefore, while this is understandable, it further complicates the use of IQ tests for such diagnoses, as it indicates that a child tested on an outdated, easier IQ test is being directly compared to a child tested on a current, harder test in the same year (and even possibly in the same classroom).

The broader implications of this research, however, raise questions about the emphasis that is placed on an IQ. Although IQs are rising and falling in a (now) predictable fashion, it is important to remember that this is occurring regardless of real cognitive changes—even the same person taking both IQ tests would score lower on the newer test. Yet these IQs are often compared against on another as though they are of the same metric (e.g., a student with a WISC-R IQ of 65 is treated the same as a student with a WISC-III IQ of 65). Psychologists and professionals who deal with and interpret IQ need to be aware of this fallacy.

Acknowledgements

This project was funded by a grant from the Smith Richardson awarded to the second author and the Cornell University College of Human Ecology Graduate Student Research Grant awarded to the first author.

References

American Association of Mental Retardation (2002). *Mental retardation: Definition, classification, and systems of supports* (10th ed.). Annapolis, MD: Author.

Bolen, L. M., Aichinger, K. S., Hall, C. W., & Webster, R. E. (1995). A comparison of the performance of cognitively disabled children on the WISC-R and WISC-III. *Journal of Clinical Psychology*, *51*, 89–94.

Buck v Bell, 274 US 200 (1927).

Burke, A. A. (1975). Placement of black and white children in educable mentally handicapped and learning disability classes. *Exceptional Children*, *41*, 438–439.

S. J., Ceci (Ed.). (1985). *Handbook of cognitive, social, and neuropsychological aspects of learning disability.* Hillsdale, NJ: Erlbaum Associates.

Edgerton, R. B. (1967). *The cloak of competence: Stigma in the lives of the mentally retarded*. Berkeley: University of California Press.

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

Flynn, J. R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal of Mental Deficiency*, *90*, 236–244.

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

Flynn, J. R (1998). IQ gains over time: Toward finding the causes. In: Neisser U. (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 3–66). Washington, DC: American Psychological Association.

Flynn, J. R. (2000). The hidden history of IQ and special education: Can the problems be solved? *Psychology, Public Policy, and Law*, *6*, 191–198.

Gaskill, F. W. III, & Brantley, J. C. (1996). Changes in ability and achievement scores over time: Implications for children classified as learning disabled. *Journal of Psychoeducational Assessment*, *14*, 220–228.

Gould, S. J. (1981). The mismeasure of man. New York, NY: W.W. Norton.

K. A., Heller, W. H. Holtzman, & S. Messick (Eds.). (1982). *Placing children in special education; A strategy for equity.* Washington, DC: National Academy Press.

Herrnstein, R. J., & Murray, C. (1994). The bell curve. New York, NY: Free Press.

Kavale, K. A., & Forness, S. R. (1992). History, definition and diagnosis. In: Singh N. N., & Beale I. L. (Eds.), *Learning disabilities: Nature, theory, and treatment*. New York, NY: Springer.

Kavale, K. A., & Forness, S. R. (2000). What definitions of learning disability say and don't say: A critical analysis. *Journal of Learning Disabilities*, *33*, 239–256.

Mercer, J. R. (1973). Labeling the mentally retarded. Berkeley, CA: University of California Press.

National Advisory Committee on Handicapped Children (1968). *First annual report: Special education for* *handicapped children*. Washington, DC: US Office of Education, Department of Health, Education and Welfare.

Public Law 94-142 (November 29, 1975). *Education for all handicapped children act. US Congress.*

N. N., Singh, & I. L., Beale (Eds.). (1992). *Learning disabilities: Nature, theory, and treatment*. New York, NY: Springer.

Slate, J. R., & Saarnio, D. A. (1995). Differences between WISC-III and WISC-R IQs: A preliminary investigation. *Journal of Psychoeducational Assessment*, *13*, 340–346.

Spitz, H. H. (1983). Intratest and intertest reliability and stability of the WISC, WISC-R, and WAIS Full Scale IQs in a mentally retarded population. *Journal of Special Education*, *17*, 69–80.

Spitz, H. H. (1989). Variations in Wechsler interscale IQ disparities at different levels of IQ. *Intelligence*, *13*, 157–167.

Stinnet, T. A., Harvey, J.M, & Oehler-Stinnet, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment*, *12*, 331–350.

Truscott, S. D., & Frank, A. J. (2001). Does the Flynn effect affect IQ scores of students classified as LD? *Journal of School Psychology*, *39*, 319–334.

US Department of Education (2001). *Twenty-third annual report to Congress on the implementation of the individuals with disabilities education act*. Washington, DC: Author.

Vance, H., Maddux, C. D., Fuller, G. B., & Awadh, A. M. (1996). A longitudinal comparison of WISC-III and WISC-R scores of special education students. *Psychology in the Schools*, *33*, 113–118.

Vitello, S. J., & Soskin, R. M. (1985). Mental retardation: Its social and legal consequences. Englewood Cliffs, NJ: Prentice-Hall.

Wechsler, D. (1974). *The Wechsler intelligence scale for children-revised manual*. New York, NY: The Psychological.

Wechsler, D. (1991). *The Wechsler intelligence scale for children-III manual*. New York, NY: The Psychological.