Intelligence 48 (2015) 85–95

Contents lists available at ScienceDirect

Intelligence

The magical numbers 7 and 4 are resistant to the Flynn effect: No evidence for increases in forward or backward recall across 85 years of data

Gilles E. Gignac

School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia, 6009, Australia

ARTICLE INFO

Article history:
Received 5 July 2014
Received in revised form 23 September 2014
Accepted 3 November 2014
Available online xxxx

Keywords:
Flynn effect
Short-term memory
Working memory
Miller's law

ABSTRACT

A substantial amount of empirical research suggests that cognitive ability test scores are increasing by approximately three IQ points per decade. The effect, referred to as the Flynn effect, has been found to be more substantial on measures of fluid intelligence, a construct known to be substantially correlated with memory span. Miller (1956) suggested that the typical short-term memory capacity (STMC) of an adult is seven, plus or minus two objects. Cowan (2005) suggested that the typical working memory capacity (WMC) of an adult is four, plus or minus one object. However, the possibility that both STMC and WMC test scores may be increasing across time, in line with the Flynn effect, does not appear to have been tested comprehensively yet. Based on Digit Span Forward (DSF) and Digit Span Backward (DSB) adult test scores across 85 years of data (respective *Ns* of 7,077 and 6,841), the mean adult verbal STMC was estimated at 6.56 (±2.39), and the mean adult verbal WMC was estimated at 4.88 (±2.58). No increasing trend in the STMC or WMC test scores was observed from 1923 to 2008, suggesting that these two cognitive processes are unaffected by the Flynn effect. Consequently, if the Flynn effect is occurring, it would appear to be a phenomenon that is completely independent of STMC and WMC, which may be surprising, given the close correspondence between WMC and fluid intelligence.

1. Introduction

One of the most sensational scientific observations in the area of contemporary intelligence research is that intelligence test scores have increased since about 1930 (Flynn, 2012; Lynn, 1982). The reported effect is not small, as it corresponds to approximately three IQ points per decade (Flynn, 2007; Neisser, 1998). Furthermore, the consequences are not negligible, as Flynn (1987) contended that the "…gains suggest that IQ tests do not measure intelligence but rather a weak causal link to intelligence" (p. 190). The precise nature and causes of the "Flynn effect" remain enigmatic (Williams, 2013). Furthermore, a number of limitations associated with studies supportive of the Flynn effect have been articulated, including invalid test score comparisons due to changes in test items and administration across editions (Kaufman, 2010), changes in the rate of human cognitive development in both the young and the elderly (Parker, 1986), changes in standard deviations (Rodgers, 1998), as well as the absence of factorial invariance associated with intelligence battery test scores Must, te Nijenhuis, Must, & van Vianen, 2009; Wicherts et al., 2004). Consequently, the purpose of this investigation was to examine the Flynn effect on several normative samples at the observed score level on possibly the only subtest of intellectual functioning that has essentially not changed for over a century: Digit Span. As Digit Span incorporates both forward and backward recall items, an additional purpose of this investigation was to estimate precisely the typical verbal short-term memory capacity (STMC) and working memory capacity (WMC) of adults, so as to verify the proposed values reported by Miller (1956; 7 ±2) and Cowan (2005; 4 ±1).

E-mail address: gilles.gignac@uwa.edu.au.

http://dx.doi.org/10.1016/j.intell.2014.11.001
0160-2896/

1.1. Overview of the Flynn Effect

The accumulated research suggests that the Flynn effect is more pronounced on fluid intelligence tests, in comparison to tests likely to be affected by education, such as vocabulary and knowledge of worldly facts (Flynn, 2007; Rönnlund, Carlstedt, Blomstedt, Nilsson, & Weinehall, 2013). In a relatively recent investigation, Flynn (2009a) reported ongoing IQ gains (1943 to 2008) in British children (5.5 to 11 years old) as measured by the Raven's Progressive Matrices (Raven, Court, & Raven, 1986; Raven, Rust, & Squire, 2008). Additionally, Flynn (2009b) reported continued (1995–2006) IQ increases equal to three IQ points per decade in adults based on the Wechsler scales. Based on an examination of the Seattle Longitudinal Study (SLS) database, Schaie, Willis, and Pennak (2005) reported a Flynn effect equal to approximately ½ of a standard deviation in cognitive ability test scores between birth cohort 1931 and birth cohort 1952. As the results were most pronounced for inductive reasoning, Schaie et al. (2005) recommended that it would be insightful to evaluate possible test score changes across time in fluid type capacities more basic than inductive reasoning.

Arguably, one such relatively elementary cognitive ability construct is memory span. Individual differences in memory span (WMC in particular) are known to be correlated substantially with fluid intelligence. Based on a meta-analysis, Kane, Hambrick, and Conway (2005) estimated that approximately 50% of the true score variance between WMC and fluid intelligence is shared. Based on the WAIS-IV normative sample, Gignac (2014) suggested that the shared variance may be closer to 60%. The substantial empirical association between WMC and fluid intelligence is considered an important phenomenon, as it has been theorised that WMC is a critical determinant, or rate limiting factor, in the performance of fluid intelligence tasks (Carpenter, Just, & Shell, 1990; Fry & Hale, 1996). Oberauer, Su, Wilhelm, and Sander (2007) proposed that the association between WMC and fluid intelligence is embedded by the central nervous system in such a way that only a limited number of bindings can be created to facilitate the development of novel relational representations. Consequently, given the close correspondence between WMC and fluid intelligence on both empirical and theoretical grounds, the reported increases in fluid intelligence test scores (Flynn, 2007) would arguably be expected to be associated with concomitant increases in memory span, particularly WMC.

1.2. The case for Digit Span

One of the most commonly used tests of memory span is Digit Span (Blankenship, 1938; Dempster, 1981). According to Bronner, Healy, Lowe, and Shimberg (1927), Digit Span was in use as early as 1887. Digit Span's popularity was established by virtue of the fact that it was included in both of the intelligence batteries that emerged as the most popular in the early 20th century: the Stanford-Binet (Terman, 1917); and the Wechsler-Bellevue scale (W-B; Wechsler, 1939). Although there are several slight variations of the Digit Span subtest, typically, the test consists of administering several series of single digits to be recalled in a particular order. In most cases, the number of digits within a series ranges from 3 to 9. There are two common forms of the Digit Span test: Digit Span Forward (DSF), where the digits need to be recalled in the order with which they were presented, and Digit Span Backward (DSB), where the digits need to be recalled in the reverse order with which they were presented.

Although Digit Span was initially considered a relatively poor measure of intellectual functioning (Matarazzo, 1972; Wechsler, 1939), such a position appears to be based more on presumption and clinical experience, rather than rigorous statistical evidence (Bachelder & Denny, 1977; Verive & McDaniel, 1996). For example, Wechsler (1939) presumed that there was not a sufficient amount of variability in Digit Span scores to be a high quality discriminator of intelligence, as approximately 90% of the adult population appeared to recall somewhere between five and eight digits. Additionally, Wechsler (1939) claimed that both DSF and DSB correlated poorly with other intelligence subtests and contained little of *g*. However, Wechsler's (1939) own reported results do not support such a position. First, based on the Wechsler-Bellevue (Wechsler, 1939) normative sample (ages: 20–34, $N = 355$), Digit Span was associated with a mean inter-subtest correlation of .38, which is comparable to the mean inter-subtest correlation of .44 for the whole battery. Additionally, based on the same portion of the normative sample, Wechsler (1939) reported the corrected subtest-FSIQ correlation (a reasonable proxy of a *g* component loading) associated with Digit Span at .51, which, arguably, was not substantially smaller than the average corrected subtest-FSIQ correlation of .61. More recently, based on the Wechsler Adult Intelligence Scale – IV (WAIS-IV; Wechsler, 2008) normative sample ($N = 2{,}200$) and a bifactor model, Gignac (2014) found that DSF and DSB were associated with *g* loadings of .46 and .58, respectively, which would suggest that both subtests are moderate indicators of *g*. Disattenuated for imperfect reliability in subtest scores, the corrected *g* loadings corresponded to .51 and .64, respectively. Jensen and Figueroa (1975) also found that DSB correlated more significantly with *g* than DSF. Thus, although Digit Span is certainly not an excellent indicator of *g*, it is arguably a fair to good indicator of intellectual functioning, particularly DSB.

Digit Span has also been observed to share variance with a number of socially important variables. For example, Frank (1983) reviewed four studies (seven independent samples) which examined the association between the Wechsler subscales and grade point average. Digit Span was associated with a mean validity coefficient of .35, which was very comparable to the mean validity coefficient of .37 across all 11 subtests. Digit Span has also been found to correlate with years of education completed ($r = .44$, Paul et al., 2005; $r = .43$, Birren & Morrison, 1961), reading comprehension ($r = .30$; Daneman & Merikle, 1996; Norman, Kemper, & Kynette, 1992), and brain volume ($r = .41$; Wickett, Vernon, & Lee, 2000). Additionally, amongst a battery of cognitive ability tests, Digit Span was found to be the best predictor of academic achievement amongst learning-problem children (Serwer, Shapiro, & Shapiro, 1972). Digit Span has also been found to be a respectable predictor of job performance (medium cognitive demands: $r = .51$; Verive & McDaniel, 1996). Finally, Miller and Vernon (1992) found that the association between reaction time and *g* was mediated by individual differences in short-term memory span. Thus, in light of the above, it is likely tenable to suggest that Digit Span is somewhere between a moderate to good indicator of