Psychological Injury and Law
https://doi.org/10.1007/s12207-024-09499-3

Measurement of Working Memory on the Wechsler Adult Intelligence Scale: Should We Subtract Arithmetic?

Allyson G. Harrison[1] · **A. Lynne Beal**[2] · **Irene T. Armstrong**[1] · **Anna Gallagher**[1]

Received: 30 November 2023 / Accepted: 19 January 2024

Abstract
Deficits in working memory are often used as evidence to support the diagnosis of a neuropsychological disorder or acquired brain injury. Subtests that comprise the working memory index (WMI) of the WAIS-IV are assumed to all measure the same construct; however, deficits in basic arithmetic skills may unduly influence performance on the Arithmetic subtest, particularly among students with neurodevelopmental disabilities. The current study examined the relationship between working memory scores and measured mathematical ability in a sample of 605 postsecondary students seeking psychoeducational assessments to verify their need for academic accommodation. Different combinations of the three WAIS-IV working memory subtests' scores resulted in significantly different composite WMI scores. Across diagnostic groups, subtest scores were highest on Letter-Number Sequencing and Digit Span and lowest on Arithmetic. Furthermore, math achievement scores were more strongly correlated with Arithmetic scores than with Digit Span or Letter-Number Sequencing scores. Students with learning disabilities performed significantly more poorly on Arithmetic relative to students with attention deficit hyperactivity disorder. Given the risk of clinicians falsely diagnosing a working memory impairment, we therefore advise clinicians to substitute out the Arithmetic subtest, which requires math skills, when evaluating young adults for neurodevelopmental conditions. This will avoid them being falsely identified as having a working memory impairment when in fact their working memory is intact.

Keywords Math performance · WAIS-IV · Working memory · Arithmetic

Introduction

Working memory (or short-term memory) has been broadly defined as the ability to temporarily store and, at times, manipulate information in the mind, so that this information can be applied to tasks (Baddeley, 2012). Cowan (2014) described working memory as the underpinning of cognitive development, learning, and academic achievement. Similarly, Schneider and McGrew (2018) have defined working memory capacity, thought to be an important precursor to all academic achievement, as the ability to maintain and manipulate information in active attention. They have further identified four narrow abilities within working memory: auditory short-term storage; visual-spatial short-term storage, attentional control; and working memory capacity.

Working memory deficits have been implicated as a causal factor for symptoms of numerous neurodevelopmental disorders, including reading disabilities (Swanson et al., 2009), math disabilities (Andersson & Lyxell, 2007; Attout & Majerus, 2015), attention deficit hyperactivity disorder (Kofler et al., 2020), and autism spectrum disorders (Habib et al., 2019). Working memory has been found to contribute to basic learning skills, such as the ability to follow instructions (Yang et al., 2016) and solve problems (Wiley & Jarosz, 2012). Working memory has also been empirically associated with the development of abilities of reading (Pham & Hasson, 2014), writing (Bourke & Adams, 2010), and math (Alloway & Passolunghi, 2011). Working memory deficits are also said to be common neurocognitive sequalae of concussions (e.g., McAllister et al., 2006) and mild traumatic brain injuries (e.g., Carroll et al., 2004).

As part of a comprehensive psychological assessment, psychologists commonly administer tests of working memory to inform conceptualizations of diagnostic presentation

✉ Allyson G. Harrison
harrisna@queensu.ca

[1] Regional Assessment and Resource Centre, Queen's University, Kingston, Canada

[2] Private Practice, Toronto, Canada

Published online: 04 February 2024

and treatment recommendations. In such cases, the accuracy of a clinician's conclusions is heavily reliant on the construct validity of the selected working memory test—that is, does the test measure what it purports to measure?

Measuring Working Memory on the Wechsler Scales

Subtests measuring working memory are integral to tests of cognitive abilities. The most recent fourth edition of the Wechsler Adult Intelligence Scale (WAIS-IV; Weschler, 2008a) is a measure of intelligence for adults aged 16 years and above. It is based on a four-factor structure of intelligence, including of particular relevance to the current study, the working memory index (WMI). The WAIS-IV WMI is purported to measure working memory (simultaneous and sequential processing), attention, and concentration (Wechsler, 2008b). Despite being hailed as the "gold standard" of psychological assessment tools (e.g., Hartman, 2009), a growing body of theoretical and empirical evidence has called into question the construct validity of the WAIS-IV WMI.

The Wechsler scales have traditionally included three subtests said to measure working memory: Digit Span (DS), Arithmetic (ARITH), and Letter-Number Sequencing (LNS). On DS, the examinee first repeats, and then later reorders, sequences of numbers. DS is purported to involve rote learning and memory, attention, encoding, and auditory processing as well as working memory, transformation of information, mental manipulation, and visuospatial imaging (Wechsler, 2008b p. 15). On ARITH, the examinee mentally solves timed arithmetic word problems. ARITH is purported to involve mental manipulation, concentration, attention, short- and long-term memory, numerical reasoning ability, and mental alertness and may also involve sequential processing, fluid, quantitative and logical reasoning, and quantitative knowledge (Wechsler, 2008b p. 15). On LNS, the examinee mentally rearranges increasingly complex sequences of numbers and letters in alphabetical and numerical order. LNS is purported to involve sequential processing, mental manipulation, attention, concentration, memory span, and short-term auditory memory (Wechsler, 2008b p. 16).

Starting with the third generation of the Wechsler scales, Wechsler Intelligence Scale for Children-III (WISC-III; Wechsler, 1991), and Wechsler Adult Intelligence Scale-III (WAIS-III; Wechsler, 1997), the publisher developed a working memory factor. On the WISC-III, it was called the Freedom From Distractibility Index, comprising DS and ARITH. On the WAIS-III, it was called the working memory index (WMI), comprising DS, ARITH, and LNS.

In the subsequent editions of the WISC, the core subtests of the WMI did not include ARITH. The WISC-IV (Wechsler, 2003) used DS and LNS, while ARITH became a supplementary measure of working memory. The WISC-V (Wechsler, 2014) eliminated ARITH from the WMI since ARITH was found to not be a strong empirical measure of short-term memory; only DS and Picture Span (PS) are now used. Factor analysis of the WISC-V normative data found ARITH to load on fluid reasoning (0.32), working memory (0.31), and verbal comprehension (0.16) (Wechsler, 2014). Weiss et al. (2013a) also established a relationship of ARITH with fluid reasoning. Additionally, Reynolds and Keith (2017) found ARITH to measure a mix of fluid reasoning, working memory, and verbal comprehension. As such, ARITH was rebranded as a measure of quantitative reasoning on the WISC-V.

In contrast, the Fourth Edition of the WAIS (WAIS-IV; Wechsler, 2008a) uses DS and ARITH as primary measures of WMI and made LNS a supplementary subtest. The WAIS-IV Administration and Scoring Manual (Wechsler, 2008c p. 29) permits substitution of the supplementary subtest LNS for either core subtest DS or ARITH to calculate WMI for ages 16 to 69, in clinical situations where when an examinee's physical condition interferes with performance, or if a core subtest is invalidated for any reason; but not for the sole purpose of changing a composite score. The test author notes that WAIS-IV ARITH items "were revised to decrease demands on verbal comprehension and mathematical knowledge, thus increasing demands on working memory" (Wechsler, 2008b, p. 18). Still, the subtest involves "mental manipulation, concentration, attention, short-and long-term memory, numerical reasoning ability, and mental alertness. It may also involve sequential processing; fluid, quantitative, and logical reasoning; and quantitative knowledge" (p. 15).

Wechsler (2008b) explained that the change of core subtests in the WAIS-IV WMI to exclude LNS followed from results of factor analytic studies of the normative sample. For ages 16 to 69, they found factor loadings on WMI for DS of 0.73, for LNS of 0.69, and 0.75 for ARITH. Notably, in the original factor analysis, Wechsler also found that verbal comprehension (Gc) loaded 0.08 on ARITH, whereas DS and LNS loaded only on WMI. In contrast, Lichtenberger and Kaufman (2009) reported a 0.77 factor loading of ARITH on fluid reasoning index (Gf) for the entire WAIS-IV sample. Further factor analytic studies have also shown ARITH on the WAIS-IV to cross-load on a verbal conceptual factor (Holdnack et al., 2011): DS and ARITH each had a corrected correlation of 0.60 with WMI and with each other, whereas LNS had an uncorrected correlation of 0.70 with WMI, 0.60 with DS, and 0.56 with ARITH for ages 16 to 69.

After publication of the WAIS-IV, two confirmatory factor analytic studies found evidence that ARITH loaded more heavily on a factor of fluid reasoning than one that involves short-term or working memory (Gsm). Benson and colleagues (2010) completed an independent confirmatory factor analysis of the WAIS-IV using a Cattell-Horn-Carroll model of intelligence (CHC; Carroll, 1993). They found that ARITH was a strong measure of fluid reasoning abilities (fluid reasoning loaded 0.54 on ARITH), as opposed to Gsm (0.28). In comparison, both DS and LNS loaded highly on Gsm (0.85 and 0.80, respectively) and on no other factors (Benson et al., 2010).

Weiss and colleagues (2013b) also found support for a five-factor solution of the WAIS-IV including quantitative reasoning (RQ) consisting of ARITH and figure weights as a narrow ability subsumed under Gf. Similar to the findings of Benson et al. (2010), these authors found that ARITH loaded minimally on Gsm (0.23) but higher on RQ (0.62), while DS and LNS continued to be strong measures of Gsm. Even with a four-factor solution, ARITH continued to load weakly on the WMI (0.36), while DS and LNS had stronger loadings (0.85 and 0.80, respectively). These authors concluded that either the original four-factor model or this five-factor model could serve as a basis for WAIS-IV score interpretation. However, Canivez and Kush (2013) have criticized Weiss's factor analyses. They note that ARITH poses a factor analytic problem since there are no similar quantitative tasks in WISC-IV or WAIS-IV to better measure a latent quantitative reasoning dimension that is part of CHC conceptualization. In addition, while they acknowledge that working memory is involved in mental arithmetic tasks, they consider ARITH "also, and perhaps, more appropriate as a task of quantitative skills, and as a fair-to-good measure of *g*" (p. 165). They propose either removing ARITH altogether or creating additional quantitative reasoning tasks in subsequent versions of the Wechsler scales.

Similarly, under the Cattell-Horn-Carroll Theory (CHC), DS and LNS on the WAIS-III and WAIS-IV have consistently been categorized as empirically strong measures of short-term memory, while ARITH has not (Flanagan & Ortiz, 2001; McGrew & Flanagan, 1998). Most recently, WAIS-IV WMI has been categorized as an empirically strong mixed measure of Gsm and fluid reasoning to quantitative reasoning (Gf:RQ) (Flanagan et al., 2013, 2018).

Overall, accumulating evidence thus appears to suggest that, although warranted, revisions to the WAIS-IV ARITH may have been insufficient in addressing the construct validity of the WMI. Whether ARITH is categorized as a mixed measure of Gsm and Gf or Gsm and Gc, or whether it is related to math skills, it has not been found to be a strong empirical measure of Gsm. This raises the question of how best to measure Gsm, especially when it is important to determine if a person has a deficit in working memory that impairs their academic performance.

Relationship Between ARITH and Mathematics

Of primary importance, working memory is considered a significant contributor to mathematical ability, consistent with the factor loading of ARITH on quantitative and logical reasoning (Wechsler, 2008a; Flanagan and colleagues, 2018). Strong correlations exist between working memory and most areas of academic performance (Mather & Wendling, 2018). Both children and adults employ working memory even for simple arithmetic problems (Cragg et al., 2017), and stronger working memory ability has been associated with better performance on tests of mathematical ability in primary school children (Friso-van den Bos et al., 2013). A meta-analysis conducted by Peng and colleagues (2016) found a significant, positive association between working memory and mathematical ability in primary school children, of a medium effect size ($r = 0.35$).

The validity of ARITH in measuring working memory may, however, have disproportionate impacts for specific diagnostic groups. Indeed, it is well-established that working memory deficits are characteristic of various neurodevelopmental disorders and learning difficulties. The association between working memory and mathematical ability was found to be stronger among individuals with math learning disorders or cognitive impairment than among their neurotypical counterparts (Peng et al., 2016). Furthermore, Haberstroh and Schulte-Körne's (2022) meta-analysis of students aged 8 to 12 years or attending grades 2 to 6 identified medium deficits in short-term working memory among individuals with math difficulties, with distinct weakness in visual-spatial short-term storage. To the extent that the working memory scores included math skills, the working memory scores were diluted by these individuals' weak math skills. Deficits in working memory storage of verbal information have also been documented among individuals with reading disabilities (Maehler & Schudart, 2016). The measures of working memory in students with reading disability are clearer than in those with math disabilities. Working memory is not contaminated by reading deficits, since reading skills are not required in measures of working memory. Attention deficit hyperactivity disorder (ADHD) has also been characterized by deficits in working memory associated with ADHD symptom presentation and impairment (e.g., Kofler et al., 2020; Sarver et al., 2015). Indeed, among children with ADHD, deficits in working memory ability have been found to predict poor performance in academic contexts, above, and beyond ADHD symptom severity (Simone et al., 2018).

Since ARITH on the WAIS-IV requires mental arithmetic computation skills, a student's competency in math computations is expected to impact their WAIS-IV WMI score. We know, for instance, that level of arithmetic fluency in young adults contributes to them completing math problem-solving tasks more quickly and effectively (Price et al., 2013). Also, "There are studies that have found math fact automaticity to be a predictor of performance on general mathematics tests" (Baker & Cuevas, 2018, p. 3). This is because a lack of math facts automaticity increases the cognitive load required to do mental arithmetic (Baker & Cuevas, 2018). As such, lack of math facts automaticity (due to lack of exposure or training) may negatively influence WMI scores if math fluency (rather than Gsm) contributes significantly to performance on the timed ARITH subtest.

Students who lack basic numerical math facts fluency will overwhelm their working memory capacity when required to complete more complex math problems (Kershaw, 2010). The 2018–2019 results on the standardized province-wide tests in Ontario Canada found 46% of grade 6 students, 44% of grade 9 students studying at the applied level (preparing for community college stream), and 84% studying at the academic level (preparing for university-stream) met the provincial math standard (i.e., were performing at expected grade level). In addition, only one-third of grade 9 students enrolled in the applied math course reported themselves to be good at math (Education Quality and Accountability Office, 2019). Similar concerns have been raised in the United States. Recently, the National Assessment of Education Progress—Mathematics Assessment found that 25% of grade 4 students and 38% of all grade 8 students failed to meet basic benchmarks in math, and that math proficiency was declining in all regions of the country and in nearly all student subgroups (National Center for Education Statistics, 2022). These statistics and general trends in mathematics skills suggest that students who have deficits in math facts fluency may inadvertently score lower on ARITH and, therefore, have lower WAIS-IV WMI scores due to a cause other than a working memory deficit, per se.

Questionable face validity of ARITH may also threaten its reliability as a measure of working memory, particularly among those with math anxiety. As described by Lichtenberger & Kaufman, 2009 (p. 138) "individuals who have struggled with math in school may become anxious when asked to respond to school-like arithmetic questions," which could interfere with performance on tasks like ARITH. Among undergraduate students, math anxiety has been demonstrated to negatively impact performance on WAIS-IV ARITH but not on DS and LNS, above and beyond the effects of gender, general test anxiety, and math performance level (Buelow & Frakey, 2013). Such evidence raises the question of how best to measure working memory, especially when it is important to determine if a person has a deficit in working memory that impairs their academic performance.

In summary, DS and LNS have been demonstrated to be strong empirical measures of short-term memory, while ARITH appears to measure cognitive abilities in addition to working memory. There is also compelling evidence that difficulties in mathematics may lead to poorer performance on measures of working memory involving math calculation, such as ARITH on WAIS-IV. Thus, the inclusion of ARITH as a primary measure of working memory could lead to underestimation of an individual's true working memory abilities, particularly if that individual has poor math skills, such as in the context of a learning disability, or if math facts fluency has not been developed.

Our Hypotheses

The current study investigates the inter-relationships of subtests measuring working memory on the WAIS-IV. We also ask whether working memory index composite scores are impacted by math calculation skills in a postsecondary population.

First, we hypothesize that across diagnostic groups, scores on ARITH will be lower than scores on DS and LNS, respectively, consistent with the difficulty many students have shown with their math skills. We further explore the extent to which differences in these three subtests might be specific to certain diagnostic groups.

Second, we hypothesize that across diagnostic groups, DS and LNS will correlate more strongly with each other than with ARITH, consistent with previous factor analytic studies.

Finally, we compare scores on DS, ARITH, and LNS with scores on tests of computation and problem solving on standardized achievement tests. We expect performance on ARITH to correlate more strongly with measures of math achievement, as compared to both DS and LNS, respectively.

Method

Participants

Data for this ethics-approved study were drawn from an archival database of community college and four-year university students in Ontario Canada. These students had completed a comprehensive psycho-educational assessment at a university-based regional assessment center between 2008 and 2022 to investigate reported problems with learning, attention, or mental health challenges (e.g., anxiety and depression). All students agreed during the informed consent process to allow their de-identified information to be included in this research database.

Our sample from the database initially included 1004 participants. Individuals who failed a performance validity test were excluded from all analyses ($n = 205$). For purity of the sample, we also excluded any participants with more than a single diagnosis ($n = 194$). Thus, the final sample, utilized across analyses, included 605 participants (59.7% female). Demographic information such as race and ethnicity were not consistently collected at the time of the assessment and, therefore, are not precise. However, an evaluation of all the referrals made to the assessment center during that time period revealed the following distribution of ethnic background: 72% White, 12% Asian, 6% Black, 3% Middle Eastern, and 7% other/not specified. Based on this, we estimate that the current sample contains a similar distribution of ethnic background. Among the final sample, the assessing psychologists had diagnosed 8.6% of participants with ADHD, 43.5% with a learning disability (LD), 22.1% with a specific mental health diagnosis (MH), and 25.8% had no diagnosis (no Dx).

Participants' ages ranged from 17 to 56 years, with an average of 21.5 years (SD = 5.7 years). The groups as defined by diagnosis differed in age, $F_{(3, 598)} = 4.57$, $p = 0.004$. The learning disability group was on average significantly younger ($M = 20.6$ years, SD = 4.8 years) than the mental health group ($M = 22.8$ years, SD = 7.1 years), but neither were different in mean age from the ADHD group ($M = 21.8$ years, SD = 5.4 years) or the no diagnosis group ($M = 21.7$ years, SD = 5.6). The groups also differed in the proportion of women, with a majority of women in the LD, MH, and No Dx groups (59.2%, 60.4%, and 66.7%, respectively), but there was a minority of women in the ADHD group (40.4%), $X^2_{(3)} = 11.28$, $p = 0.010$. Age and sex did not interact across groups.

Materials

All students completed the WAIS-IV ($n = 605$), including DS, ARITH, and LNS subtests. A subset of students also completed subtests of math achievement from the Woodcock-Johnson III (WJ III, Woodcock et al., 2001; Calculation, $n = 132$ and Applied Problems, $n = 83$); Woodcock-Johnson IV (WJ IV, Schrank et al., 2014; Calculation, $n = 105$ and Applied Problems, $n = 106$); or the Wechsler Individual Achievement Test-III (WIAT III, Wechsler, 2009; Numerical Operations, $n = 183$ and math problem solving, $n = 178$). Clinicians administered at least two performance validity tests (PVTs) in each assessment, chosen from Word Memory Test (WMT; Green, 2003 revised 2005), the Medical Symptom Validity Test (MSVT; Green, 2004), Test of Memory Malingering (TOMM; Tombaugh, 1996), and the Victoria Symptom Validity Test (VSVT; Slick et al., 1997). To ensure that the current sample reflected the performance of students who were not objectively attempting to magnify or exaggerate their current symptoms, students who failed one or more PVT (based on the test's published cut-off score) were excluded.

Procedure

Students registered at either a Community College or four-year university were provided with a psycho-educational assessment at a government-funded regional assessment center. Referrals were for either an updated or new assessment to investigate currently reported academic difficulties. Students were assessed by either a registered psychologist or a supervised graduate student. The psychologists had given diagnoses where applicable, based on criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders. Although administered in Canada, test scores were calculated using the American norms to facilitate cross-battery comparison. This study was not preregistered.

Results

Repeated measures ANOVAs were conducted to analyze differences in mean performance on the three subtests of the WMI (DS, ARITH, and LNS) as a function of diagnostic group (learning disability (LD), ADHD, mental health disorder (mental health; MH), and no diagnosis (no Dx)) for WAIS-IV (see Table 1 for means and standard deviations).

Table 1 Mean and SD for scaled scores of the subtests for the WAIS-IV working memory index as a function of diagnostic group

	LD		ADHD		Mental health		No diagnosis		Total	
WAIS-IV (*n*)	263		52		134		156		605	
Subtest	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Digit span	7.6	2.0	9.1	2.1	8.9	2.3	8.4	2.8	8.2	2.4
Arithmetic	7.0	2.1	9.2	3.2	8.8	2.8	8.3	3.1	7.9	2.8
Letter-Number Sequencing	8.4	2.1	9.5	2.2	9.5	2.4	9.2	2.2	8.9	2.3
Total	7.7	2.1	9.3	2.5	9.1	2.5	8.6	2.7	8.3	2.5

Table 2 Mean and SD for WAIS-IV working memory index resulting from three combinations of subtests as a function of diagnostic group

	LD		ADHD		Mental health		No diagnosis	
WAIS-IV (*n*)	262		53		131		157	
Combination	Mean	SD	Mean	SD	Mean	SD	Mean	SD
DS/Arithmetic	84.5	9.4	95.1	11.7	93.2	12.9	90.4	14.3
DS/LNS	88.5	10.1	96.0	10.2	95.5	12.0	93.1	12.9
Arithmetic/LNS	86.9	9.1	96.2	13.0	95.0	13.2	92.7	12.8

WAIS-IV Wechsler Adult Intelligence Scale-IV, *DS* Digit Span, *LNS* Letter-Number Sequencing, *LD* learning disability, *ADHD* attention deficit hyperactivity disorder; *Mental Health* mental health disorder

There was a main effect across subtest scores such that mean subtest scores were higher for LNS than for either DS or ARITH, $F(_{2,\,601})=23.4$, $p<0.001$, Cohen's $d=0.487$ (sphericity not assumed; lower bound); however, this difference was not dependent upon group membership, such that the factors (Subtest X Group) did not interact, $F(_{3,\,601})=2.13$, $p=0.095$. Performance also depended upon group membership. The mean scores were lowest for the LD group and highest for the ADHD group, $F(_{3,\,601})=22.79$, $p<0.001$, Cohen's $d=0.904$. The ADHD group mean scores were not different from the MH or no Dx groups.

Next, we recalculated the WMI for each case in three different ways: the standard (DS and ARTH), then DS with LNS, and finally ARITH with LNS. Table 2 shows the resulting WMI as a function of each combination and by diagnostic group.

A repeated measures ANOVA with WMI combination as the repeated factor and diagnostic group as the independent factor showed a main effect of WMI combination such that WMI mean scores depend upon which combination of subtests is used, $F(_{1,\,599})=26.9$, $p<0.001$, Cohen's $d=0.518$ (sphericity not assumed). Contrast effects showed both linear and quadratic relationships such that across diagnostic groups, the DS with ARITH subtest combination results in the lowest WMI mean scores, whereas the DS with LNS combination results in the highest WMI mean scores. There was no interaction between the (Combination X Group) factors, $F(_{3,\,599})=1.9$, $p=0.130$.

Table 3 Relationship between WAIS-IV Arithmetic subtest score and math achievement subtest scores for high (> 85 or > 7) and low (≤ 85 or ≤ 7) achievers

		WAIS-IV Arithmetic				
		Low	High	X^2	*p*	Phi
WJ IV		% of total				
Calculation (*n* = 105)	Low	33.3	27.3	19.1	< 0.001	0.426
	High	4.8	34.3			
Applied Problems (*n* = 106)	Low	20.8	7.5	19.9	< 0.001	0.433
	High	18.9	52.8			
WJ III		% of total				
Calculation (*n* = 132)	Low	44.7	24.2	25.8	< 0.001	0.442
	High	5.3	25.8			
Applied Problems (*n* = 83)	Low	31.3	7.2	12.9	< 0.001	0.394
	High	25.3	36.1			
WIAT-III		% of total				
Numerical Operations (*n* = 183)	Low	31.1	18.0	40.6	< 0.001	0.471
	High	8.7	42.1			
Problem Solving (*n* = 178)	Low	25.8	6.7	52.2	< 0.001	0.542
	High	15.2	52.2			
WAIS-IV		% of total				
Digit Span (*n* = 799)	Low	13.6	10.1	42.35	< 0.001	0.230
	High	23.8	52.4			
Letter-Number Sequencing (*n* = 799)	Low	7.4	2.9	46.52	< 0.001	0.241
	High	30.0	59.7			

Performance on WAIS-IV subtests is defined as poor if the scaled score is ≤ 7; performance on WJ III and WJ IV is defined as poor if the scaled score is ≤ 85. Only 6 cases were given both WAIS-IV Arithmetic and WIAT-II

Table 4 Correlation between WAIS-IV Digit Span, Arithmetic, and Letter-Number Sequencing subtest scores and math achievement subtest scores

	WJ IV Calculation	WJ IV Applied Problems	WJ III Calculation	WJ III Applied Problems	WIAT-III Numerical Operations	WIAT-III Problem Solving
DS	0.456***	0.522***	0.410***	0.365***	0.322***	0.322***
ARITH	0.707***	0.675***	0.681***	0.636***	0.631***	0.689***
LNS	0.499***	0.560***	0.420***	0.409***	0.274***	0.357***

WAIS subtests are as follows: *DS* Digit Span, *ARITH* Arithmetic, *LNS* Letter-Number Sequencing, *WAIS* Wechsler Adult Intelligence Scale, *WIAT* Wechsler Individual Achievement Test; *WJ*, Woodcock-Johnson

***= $p < 0.001$ **= $p < 0.01$

However, there was a main effect of diagnostic group, $F_{(3, 599)} = 22.8$, $p < 0.001$, Cohen's $d = 0.716$. Post hoc pairwise comparisons showed that the learning disability group's WMI mean score was lower than the remaining three groups regardless of combination used.

To determine whether performance on WAIS-IV ARITH was related to performance in math computations or problem-solving subtests, we compared the proportions of high and low achievers, one standard deviation above or below the mean, on each subtest. Table 3 shows the results of chi-squared analysis between students scoring at or below a scaled score of 7 versus above 7 on ARITH and scoring at or below a standard score of 85 versus above 85 on tests of computation and problem solving on WJ III, WJ IV, or WIAT-III. Note that Phi coefficient values of 0.1 are considered small effects, 0.3 is a medium effect, and 0.5 is a large effect (Cohen, 1988). Results showed significant consistencies across all of the subtests in computations and problem solving: Significantly more low scorers on WAIS-IV ARITH were also low scorers on subtests of math computations and problem solving. Indeed, weak mathematical skills (low calculation and/or problem solving) are associated mainly with low scores on ARITH, and average or better math skills are associated with average or better ARITH scores. Effect sizes were medium to large. Notably, no such association was found between low scores on DS or LNS relative to low or high scores on ARITH (see Table 3); a low score on DS or LNS was not significantly predictive of how one would perform on ARITH. By contrast, although the effect was SMALL, a higher than expected number of students achieved average or better scores on both ARITH and either DS or LNS.

To determine the relative variance of each of the WAIS-IV working memory subtests related to the math achievement subtests, we computed their correlations. Table 4 shows that all correlations were significant. The highest correlations, however, were between the math achievement tests and ARITH. Finally, Table 5 shows that the correlations between DS and LNS are stronger than the correlations between ARITH and DS or between ARITH and LNS.

Table 5 Correlations among WAIS-IV WMI subscales

	Digit Span	Letter-Number Sequence
Arithmetic	0.416***	0.428***
Letter-Number Sequence	0.607***	

***= $p < 0.001$

Discussion

This study evaluated the construct validity of subtests purported to measure working memory included within the most widely used test of adult intelligence, the WAIS-IV. In a referred population of postsecondary students, our primary aim was to determine whether the WMI composite scores on the WAIS are unduly diminished by performance on ARITH, and whether performance would differ as a function of diagnostic group. Our secondary aim was to determine whether WMI composite scores are impacted by math calculation skills. First, we anticipated that DS and LNS would be stronger measures of working memory than ARITH, and that weak math calculation skills would adversely affect obtained ARITH scores, thus reducing the overall WMI score obtained. In addition, we investigated whether use of ARITH differentially affected the overall WMI scores obtained by students in any specific diagnostic group.

For WAIS-IV students consistently returned greater, mean subtest scores for both DS and LNS than for ARITH had a higher WMI score when using DS and LNS than for any WMI score calculated by including ARITH. This finding supports studies showing that WAIS-IV ARITH measures something different than do DS and LNS. The tasks of the DS and LNS subtests are consistent with Carroll's (1993) view of memory span involving the registration of the stimuli and of the order of the stimuli. ARITH has different task requirements other than short-term memory tests, including language comprehension, nonverbal reasoning, and performing arithmetic ability.

Obtained subtest scores for the WAIS-IV also varied by diagnostic group. LNS was the highest score for all diagnostic groups, whereas ARITH was the lowest score for those diagnosed with a learning disability. This finding underscores the importance of not including math-based tests when evaluating the working memory abilities of those with possible learning disabilities. However, this study did not investigate differential impacts on students with math disabilities versus reading disabilities. That question provides a direction for further research. We also did not evaluate how much math education participants had; high school students in Ontario require only 3 math credits in total in order to graduate, and these do not have to be at the academic (university-bound) level. Interestingly, students diagnosed with ADHD did best of all of the groups on all three subtests, and students with learning disabilities did significantly worse on all three subtests than did students in the other three groups.

Academically, scores obtained by our referred postsecondary sample on WAIS-IV ARITH show a consistent pattern with their scores on subtests of math computations and problem solving. Correlations between the WAIS-IV subtest scores and the math achievement subtest scores showed that ARITH accounted for more of the variance in results on math achievement tests than did either DS or LNS. This finding suggests that the score obtained on WAIS-IV ARITH by our referred postsecondary students appears to be influenced by the students' mathematics knowledge. These results are concerning. We know that many Ontario students have difficulty with math and see themselves as not being good at math (Education Quality and Accountability Office, 2019), and that 38% of all grade 8 students in the United States have deficient levels of basic math knowledge (National Center for Education Statistics, 2022). As a result, many of these students might be incorrectly identified by the WAIS-IV as having a working memory deficit when, in fact, their problem is deficient arithmetic skills.

Taken together, our results indicate that WAIS-IV ARITH is not a strong measure of working memory, at least in assessment-seeking young adults. Scores on ARITH align well with scores on tests of computation and math problem solving, showing the inter-dependence between ARITH and math skills. Our results are consistent with findings of a meta-analysis of clinical criteria for math difficulties (Haberstroh & Schulte-Korne, 2022). They found a cognitive profile of math difficulties based on criteria given by DSM-5 and ICD-10, characterized by distinct weaknesses in (a) computational calculation (calculation and fact retrieval), (b) number sense (quantity processing, quantity-number linking, and numerical relations), and (c) visual short-term memory storage. Medium weaknesses in short-term working memory were evident in persons with math disability, just as they have been found in studies of persons with reading disabilities or both math and reading disabilities (Haberstroh & Schulte-Korne, 2022). Consistently, our study found the lowest scores on ARITH for students with learning disabilities.

ARITH on the WAIS-IV correlated more highly with the scores on tests of math computations and problem solving than did DS or LNS. Our results confirm that calculating the WAIS-IV WMI based on DS and ARITH provides a measure of working memory that mixes in other abilities to a significant extent. As such, the WMI score on the WAIS-IV may not accurately evaluate actual working memory capacity. Indeed, the combination of DS and LNS consistently generated higher WMI estimates than any working memory subtest combination that included ARITH. Clinically, this results in the LD group going from having an overall WMI score falling within the impaired range when using ARITH to one that is within the statistical average range when ARITH is not included. This relationship between math skills and ARITH provides a good argument for excluding tasks involving math computations when measuring working memory. Making the permitted substitution of LNS for ARITH when calculating the WMI eliminates the additional variance inherent in ARITH and results consistently in a higher score. Diagnostically, it is especially important to get an independent measure of working memory when assessing students with poor math skills.

Our results and recommendations are consistent with the restructuring of the WISC-V that moved ARITH out of the working memory cluster to a measure of quantitative reasoning. It has been argued that ARITH may be a strong measure of overall cognitive ability, given the fact that it requires various cognitive abilities and processes (Reynolds & Keith, 2017). We agree with their recommendation that ARITH should be retained as a contributing subtest to the Full-Scale Intelligence Quotient (FSIQ), but not as a main subtest measuring auditory working memory.

If math skills are not differentiated from working memory during forensic evaluations after a reported mild brain injury, or when testing postsecondary students for disability-related accommodations, young adults might be falsely identified as having a working memory impairment when in fact their working memory is intact. Especially in disability assessments of postsecondary students, students with deficient math skills may be incorrectly identified as having working memory impairments if the WAIS-IV ARITH subtest is used to evaluate their overall working memory ability. History of math education and current performance on tests of basic arithmetic skills must be taken into consideration before concluding that working memory is impaired based on the core WAIS-IV subtests.

Like all studies, ours also has limitations. The first is that this study is of a referred sample of adult students who were seeking an assessment to determine if they could obtain accommodations in their postsecondary program. However, by eliminating data from students who failed one or more performance validity tests, we have increased confidence that the test results are valid and reliable indicators of the students' abilities. While some students were found to have no diagnosis, our sample is nonetheless under-representative of neurotypical students in this age range. Also, the age range consists mainly of Canadian students aged 18 to 25. Future studies should evaluate performance on all three WMI subtests in otherwise normal adults as well as those with neurodevelopmental or mental health diagnoses.

A further limitation of this study is that all WAIS, WIAT, and WJ tests were scored with American norms, leaving a question of generalizability of results to tests scored with norms from other countries. There are notable differences reported in the WAIS-IV normative samples between U.S. and Canadian participants (Wechsler, 2008c), and it has been established empirically that the use of Canadian norms yields significantly lower estimates of ability in this age group relative to American norms on the WAIS-IV (see Harrison et al., 2014). Thus, replication of our study in other countries using their own norms for the Wechsler scales is advisable. A final limitation is that students presenting for evaluation of post-concussion-related problems were not included in our sample. Given that working memory deficits are often used as evidence of an acquired brain injury, this study should be replicated with young adults presenting for such assessments.

In conclusion, our findings contribute to an accumulating body of research suggesting that WAIS-IV ARITH may measure cognitive abilities external to working memory, such as mathematical and verbal skills. In the context of psychoeducational assessment, the construct validity of working memory indices is especially pertinent to inform case conceptualization, diagnosis, and treatment recommendations for individuals with known or suspected neurodevelopmental disorders. Specifically, for individuals with learning difficulties, the inclusion of ARITH may lead to an underestimation of an individual's true working memory abilities.

Based on these results and the current literature, we recommend that clinicians routinely administer the supplementary LNS for WAIS-IV. Whether or not the score on ARITH is significantly lower than scores on DS and LNS, we recommend that clinicians always administer LNS and substitute LNS for ARITH to derive the WMI for postsecondary students being assessed for a possible neurodevelopmental disorder. This approach keeps ARITH in the calculation of FSIQ since it has been found to have a strong correlation with *g*, but removes its influence in the calculation of working memory. While this approach is inconsistent with the rules in the WAIS-IV Manual, it could be referenced as a footnote in the report. We advise against reporting two versions of the WMI, since that would give the impression of the examiner choosing the version that supported their hypothesis. It could also invite unqualified readers to choose the version that they preferred.

The WAIS-V (with US normative data) is set to become available sometime in 2024. While the authors have no involvement in, or inside information concerning its development, it is likely that many clinicians (including those in Canada) will continue to use and interpret scores obtained from the WAIS-IV for at least a few more years. As such, our findings should be considered by those who continue to use the current version of this test. Future studies that employ archival WAIS-IV data to evaluate questions related to working memory may also wish to consider our findings, as the results of this study may influence their choice of subtests to include for analysis. Finally, if ARITH is not included as a measure of WMI in the new WAIS-V, this study will provide users with further empirical validation supporting this decision.

Data Availability Data used in this study may be obtained by contacting the first author at harrisna@queensu.ca.

Declarations

Competing Interest The authors declare no competing interests.

References

Alloway, T. P., & Passolunghi, M. C. (2011). The relationship between working memory, IQ, and mathematical skills in children. *Learning and Individual Differences, 21*(1), 133–137. https://doi.org/10.1016/j.lindif.2010.09.013

Andersson, U., & Lyxell, B. (2007). Working memory deficit in children with mathematical difficulties: A general or specific deficit? *Journal of Experimental Child Psychology, 96*(3), 197–228. https://doi.org/10.1016/j.jecp.2006.10.001

Attout, L., & Majerus, S. (2015). Working memory deficits in developmental dyscalculia: The importance of serial order. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence, 21*(4), 432–450. https://doi.org/10.1080/09297049.2014.922170

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baker, A., & Cuevas, J. (2018). The importance of automaticity development in mathematics. *Georgia Educational Researcher*. *14(2),* Article 2. https://doi.org/10.20429/ger.2018.140202

Benson, N., Hulac, D. M. & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*(1), 121–130.

Bourke, L., & Adams, A. (2010). Cognitive constraints and the early leaning goals in writing. *Journal of Research in Reading, 33*(1). https://doi.org/10.1111/j.1467-9817.2009.01434.x

Buelow, M. T., & Frakey, L. L. (2013). Math anxiety differentially affects WAIS-IV arithmetic performance in undergraduates. *Archives of Clinical Neuropsychology: THe Official Journal of the National Academy of Neuropsychologists, 28*(4), 356–362. https://doi.org/10.1093/arclin/act006

Canivez, G. L & Kush, J. C. (2013). WAIS-IV and WISC-IV structural validity: Alternate methods, alternative results. Commentary on Weiss et al. (2013a) and Weiss et al. (2013b). *Journal of Psychoeducational Assessment*, *31*(2). https://doi.org/10.1177/0734282913478036

Carroll, L. J., Cassidy, J. D., Peloso, P. M., Borg, J., von Holst, H., Holm, L., Paniak, C., Pépin, M., & WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury (2004). Prognosis for mild traumatic brain injury: Results of the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *Journal of Rehabilitation Medicine*, (43 Suppl), 84–105. https://doi.org/10.1080/16501960410023859

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review, 26*(2), 197–223. https://doi.org/10.1007/s10648-013-9246-y

Cragg, L., Richardson, S., Hubber, P. J., Keeble, S., & Gilmore, C. (2017). When is working memory important for arithmetic? The impact of strategy and age. *PLoS ONE, 12*(12), e0188693. https://doi.org/10.1371/journal.pone.0188693

Education Quality and Accountability Office. (2019). *EQAO encourages evidence-informed discussions about math and literacy trends. Agency releases provincial-level data and contextual information*. https://www.eqao.com/en/about_eqao/media_room/Pages/2019-highlighted-provincial-results.aspx

Flanagan, D. P. & Ortiz, S. (2001). *Essentials of cross-battery assessment*. New York John Wiley & Sons, Inc.

Flanagan, D. P., Costa, M., Palma, K., Leahy, M. A., Alfonso, V. C., & Ortiz, S. O. (2018). Cross-battery assessment, the cross-battery assessment software system, and the assessment-intervention connection. In Flanagan, D. P. & McDonough, E. M. (Eds). *Contemporary intellectual assessment: Theories, tests and issues.* (731–776). Guilford.

Flanagan, D. P., Ortiz, S., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). John Wiley & Sons Inc.

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review, 10*, 29–44. https://doi.org/10.1016/j.edurev.2013.05.003

Green, P. (2004). *Green's Medical Symptom Validity Test (MSVT) for Microsoft windows: User's manual*. Green's Publishing.

Green, P. (2005). *Green's Word Memory Test*. Green's Publishing Inc.

Haberstroh, S., & Schulte-Körne, G. (2022). The cognitive profile of math difficulties: A meta-analysis based on clinical criteria. *Frontiers in Psychology, 13*, 1–22. https://doi.org/10.3389/fpsyg.2022.842391

Habib, A., Harris, L., Pollick, F., & Melville, C. (2019). A meta-analysis of working memory in individuals with autism spectrum disorders. *PLoS ONE, 14*(4), e0216198. https://doi.org/10.1371/journal.pone.0216198

Harrison, A. G., Armstrong, I. T., Harrison, L. E., Lange, R. T., & Iverson, G. L. (2014). Comparing Canadian and American normative scores on the Wechsler adult intelligence scale-Fourth Edition. *Archives of Clinical Neuropsychology, 29*, 737–746.

Hartman, D. E. (2009). Weschler Adult Intelligence Scale (WAIS-IV): Return of the gold-standard. *Applied Neuropsychology, 16*(1), 85–97. https://doi.org/10.1080/09084280802644466

Holdnack, J. A., Zhou, X., Larrabee, G. J., Mills, S. R., & Salthouse, T. A. (2011). Confirmatory factor analysis of the WAIS-IV/WMS-IV. *Assessment, 18*(2), 178–191. https://doi.org/10.1177/1073191110393106

Kershaw, A. (2010). Big drop in math skills of entering students. *University Affairs*, Retrieved from: http://www.universityaffairs.ca/big-drop-in-math-skills-of-entering-students.aspx

Kofler, M. J., Singh, L. J., Soto, E. F., Chan, E. S. M., Miller, C. E., Harmon, S. L., & Spiegel, J. A. (2020). Working memory and short-term memory deficits in ADHD: A bifactor modeling approach. *Neuropsychology, 34*(6), 686–698. https://doi.org/10.1037/neu0000641

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essentials of WAIS-IV assessment*. Wiley.

Maehler, C., & Schuchardt, K. (2016). Working memory in children with specific learning disorders and/or attention deficits. *Learning and Individual Differences, 49*, 341–347. https://doi.org/10.1016/j.lindif.2016.05.007

Mather, N., & Wendling, B. (2018) Linking cognitive abilities to academic interventions for students with specific learning disabilities. In Flanagan, D.P. & McDonough, E.M. (Eds). *Contemporary intellectual assessment: Theories, tests and issues.* (731–776). Guilford.

McAllister, T. W., Flashman, L. A., McDonald, B. C., & Saykin, A. J. (2006). Mechanisms of working memory dysfunction after mild and moderate TBI: Evidence from functional MRI and neurogenetics. *Journal of Neurotrauma, 23*(10), 1450–1467. https://doi.org/10.1089/neu.2006.23.1450

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR) Gf-Gc cross-battery assessment*. Allyn & Bacon.

National Center for Education Statistics (2022). *National Assessment of Education Progress (NAEP) report card: 2022 NAEP Mathematics Assessment*. Retrieved from: https://www.nationsreportcard.gov/mathematics/?grade=4

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology, 108*(4), 455–473. https://doi.org/10.1037/edu0000079

Pham, A. V., & Hasson, R. M. (2014). Verbal and visuospatial working memory as predictors of children's reading ability. *Archives of Clinical Neuropsychology, 29*(5), 467–477. https://doi.org/10.1093/arclin/acu024

Price, G. R., Mazzocco, M. M., & Ansari, D. (2013). Why mental arithmetic counts: Brain activation during single digit arithmetic predicts high-school math scores. *Journal of Neuroscience, 33*, 156–163.

Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children – Fifth Edition: What does it measure? *Intelligence, 62*, 31–47.

Sarver, D. E., Rapport, M. D., Kofler, M. J., Raiker, J. S., & Friedman, L. M. (2015). Hyperactivity in attention-deficit/hyperactivity disorder (ADHD): Impairing deficit or compensatory behavior? *Journal of Abnormal Child Psychology, 43*, 1219–1232.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In Flanagan, D. P. & McDonough,

E. M. *Contemporary Intellectual Assessment Theories, Tests and Issues Fourth Edition*. Guilford.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV tests of achievement*. Rolling Meadows, IL: Riverside.

Simone, A. N., Marks, D. J., Bédard, A. C., & Halperin, J. M. (2018). Low working memory rather than ADHD symptoms predicts poor academic achievement in school-aged children. *Journal of Abnormal Child Psychology, 46*, 277–290.

Slick, D., Hopp, G., Strauss, E., & Thompson, G. B. (1997). Victoria symptom validity test, version 1.0 professional manual. *Psychological Assessment Resources.*

Swanson, H. L., Zheng, X., & Jerman, O. (2009). Working memory, short-term memory, and reading disabilities: A selective meta-analysis of the literature. *Journal of Learning Disabilities, 42*(3), 260–287. https://doi.org/10.1177/0022219409331958

Tombaugh, T. N. (1996). *Test of memory malingering TOMM*. Multi-Health Systems Inc.

Wechsler, D. (1997). *Wechsler adult intelligence scale-III.* San Antonio, Texas: Pearson.

Wechsler, D. (2008a). *Wechsler adult intelligence scale-IV.* San Antonio, Texas: Pearson.

Wechsler, D. (2008b). *WAIS-IV technical and interpretive manual.* San Antonio, Texas: Pearson.

Wechsler, D. (2008c). *WAIS-IV administration and scoring manual.* San Antonio, Texas: Pearson.

Wechsler, D. (2009). *Wechsler Individual Achievement Test – 3rd ed.* Pearson.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children - Third edition (WISC-III)*. The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* – (4th ed.). Psychological Corporation.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* – (5th ed.). Pearson.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013a). WISC-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31*(2), 114–131. https://doi.org/10.1177/0734282913478032

Weiss, L. G., Keith, T. Z. & Chen, H. (2013b). WAIS-IV and clinical validation of the four- and five-factor interpretive approaches. *Journal of Psychoeducational Assessment*, *31*(2). https://doi.org/10.1177/0734282913478030

Wiley, J., & Jarosz, A. F. (2012). "How working memory capacity affects problem solving," in. B. H. Ross (Ed). *Psychology of learning and motivation* (Elsevier Inc), 185–227. https://doi.org/10.1016/B978-0-12-394393-4.00006-6

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside.

Yang, T., Allen, R. J., & Gathercole, S. E. (2016). Examining the role of working memory resources in following spoken instructions. *Journal of Cognitive Psychology, 28*(2), 186–198. https://doi.org/10.1080/20445911.2015.1101118