Effects of Practice on the Wechsler Adult Intelligence Scale-IV Across 3- and 6-Month Intervals

Eduardo Estevis¹, Michael R. Basso¹, and Dennis Combs²

¹Department of Psychology, University of Tulsa, Tulsa, OK, USA
²Department of Psychology and Counseling, University of Texas at Tyler, Tyler, TX, USA

A total of 54 participants (age \(M = 20.9\); education \(M = 14.9\); initial Full Scale IQ \(M = 111.6\)) were administered the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) at baseline and again either 3 or 6 months later. Scores on the Full Scale IQ, Verbal Comprehension, Working Memory, Perceptual Reasoning, Processing Speed, and General Ability Indices improved approximately 7, 5, 4, 5, 9, and 6 points, respectively, and increases were similar regardless of whether the re-examination occurred over 3- or 6-month intervals. Reliable change indices (RCI) were computed using the simple difference and bivariate regression methods, providing estimated base rates of change across time. The regression method provided more accurate estimates of reliable change than did the simple difference between baseline and follow-up scores. These findings suggest that prior exposure to the WAIS-IV results in significant score increments. These gains reflect practice effects instead of genuine intellectual changes, which may lead to errors in clinical judgment.

Keywords: Neuropsychological assessments; Inpatient; Depression.

INTRODUCTION

Neuropsychological tests are frequently re-administered in clinical or forensic settings. In particular, serial neuropsychological assessments are administered to monitor the progression of a neurological disorder, to evaluate the therapeutic efficacy of a drug, or to monitor the effectiveness of a rehabilitation program (McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992). Furthermore, in a forensic setting, serial neuropsychological assessments are conducted in medico-legal cases in order to assess the plaintiff’s level of functioning and course of recovery (McCaffrey & Westervelt, 1995). Intellectual assessments are also used in the judicial system in order to assess a defendant’s competence to stand trial, criminal responsibility, Miranda comprehension and appreciation, mitigating issues in sentencing, and mental retardation (Frumkin, 2006). In this context, repeated neuropsychological assessments may occur during extended appeals or litigation, and courts may order up to four independent neuropsychological assessments within 68–90 days (Duvall & Morris, 2006).

Despite the importance of serial neuropsychological assessments, interpretation of functional change may be obscured by extraneous influences, such as...
practice effects. Practice effects reflect change resulting from repeated administration of neuropsychological tests. This occurs when a patient’s neuropsychological performance on an assessment improves due to the re-administration of the same testing instrument (McCaffrey & Westervelt, 1995). There is a strong propensity for practice effects to bias test scores under the following conditions: (1) factors that are specific to a certain assessment are memorized and (2) instructions of test procedures are memorized (Rapport, Brines, Axelrod, & Theisen, 1997). Although brain damage may result in memory impairment, practice effects have been reported in both neurologically normal individuals and in patients with neurological disease (McCaffrey, Duff, & Westervelt, 2000a, 2000b). Because cognitive functioning among normal adults tends to remain relatively stable across short intervals such as months, improvements on serial assessments may be attributed, at least in part, to practice effects rather than genuine change (McCaffrey & Westervelt, 1995).

Test characteristics may further moderate the presence of practice effects. In particular, assessments that assess speed of response require the use of an infrequent practice response, or involve a distinct and memorable solution are susceptible to practice effects (Basso, Bornstein, & Lang, 1999; Dodrill & Troupin, 1975). Practice effects occur on serial assessments because the individual under examination learns to respond more effectively after the initial examination (Lezak, 2004).

The Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV) (Wechsler, 2008) is commonly used as part of neuropsychological assessments. Presently, little research has examined the magnitude of test improvement on the WAIS-IV across time. However, investigations involving the Wechsler Adult Intelligence Scale-Revised (WAIS-R) revealed an increase in Full Scale IQ (FSIQ), Performance IQ (PIQ), and Verbal IQ (VIQ) scores when individuals were retested with this instrument (Kaufman & Lichtenberger, 1999; Wechsler, 1981). These increases occurred in both normal and clinical samples (Kaufman, 1990), and were present over a period as long as 8 months (Moore et al., 1990). More recently Basso, Carona, Lowery, and Axelrod (2002) re-administered the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) to a group of healthy control participants after either 3- or 6-month intervals. Regardless of retest interval, FSIQ, Verbal Comprehension Index (VCI), Perceptual Organization Index (POI), and Processing Speed Index (PSI) scores increased significantly upon re-examination. The magnitude of test gains was equivalent between the two retest intervals. Moreover, the magnitude of test score gains paralleled those reported in the WAIS-III technical manual. In that study 394 individuals were re-administered the WAIS-III over 1 month.

With respect to the WAIS-IV, its technical manual reports a study in which 298 individuals were administered the battery on two occasions. The interval of re-administration ranged from 8 to 82 days with a mean interval of 22 days (Coalson & Raiford, 2008). These individuals were classified across four age bands, including: 16–29, 30–54, 55–69, and 70–90. Each group displayed significant increases on all subtests. Additionally, scores on the Perceptual Reasoning (PRI) and PSI subtests increased more than the VCI and Working Memory Index (WMI) subtests. Specifically, scores increased from 2.1–3.4 points for VCI, 3.8–5.0 points for PRI, 2.0–4.6 points for WMI, and 3.2–5.6 points for PSI. Overall, FSIQ
scores improved from 3.9–5.0 points. These findings parallel those reported in the WAIS-III technical manual concerning a similar study of the WAIS-IV’s predecessor (Tulsky & Zhu, 1997), and they are comparable to the results reported by Basso et al. (2002) and Moore et al. (1990).

Despite the thoroughness of the study reported in the WAIS-IV technical manual, the average retest interval was only 22 days (Coalson & Raiford, 2008). Such an interval may lack practical application to clinicians and researchers who tend to re-assess neuropsychological functioning over longer intervals such as 3, 6, or 12 months (e.g., Basso et al., 2002). Therefore it would be important to determine the magnitude of practice effects on the WAIS-IV over such intervals. Elaborating on this question, WAIS-IV score increases may vary across 3- or 6-month intervals. Inasmuch as memory for the assessment procedures may diminish over time, smaller score increments may occur over longer intervals. Notably, Basso et al. (2002) failed to find a moderating effect of inter-test interval on WAIS-III increments. Yet the WAIS-IV has been substantially revised, and inter-test interval may moderate the effect of time with the new version of the instrument.

The purpose of this study is to assess the effects of practice on WAIS-IV performance as a result of prior testing, and whether the effects of practice vary across 3- or 6-month testing intervals. To assess this, participants were administered the WAIS-IV at baseline and again, either 3 or 6 months later. Participants were healthy individuals, thereby ensuring that score fluctuations were due to practice and random test error rather than actual improvement in function associated with recovery from cerebral dysfunction.

Although mean differences across time reveal how much change may be expected due to practice, such data fail to clarify whether fluctuations for an individual are statistically significant. Consequently, reliable change indices (RCI) were computed. RCIs establish what amount of change is expected due to test error. Score variations that exceed these estimated base-rates of change are presumed to reflect clinically meaningful improvement or decline. Several methods of estimating RCIs have been proposed. In one approach (e.g., Chelune, Naugle, Luders, Sedlak, & Awad, 1993), a confidence interval is computed using the standard error of the difference (SED). If the difference between actual retest and baseline test scores exceeds this confidence interval, clinically meaningful change has presumably occurred. Another method of computing RCIs utilizes the standard error of prediction (SEP), and evaluates differences between predicted and actual scores during the retest interval (Charter, 2009; Heaton et al., 2001; Temkin, Heaton, Grant, & Dikmen, 1999). With this method, baseline scores are predictors in regression equations, and retest scores are predicted. The SEP is used to compute a confidence interval. If the difference between the obtained and the predicted retest scores exceeds this confidence interval, clinically meaningful change is inferred to have occurred. Studies have compared the validity of these methods, and have generally found that regression-based methods employing the SEP are more accurate than simple difference methods employing the SED (Heaton et al., 2001; Temkin et al., 1999). However, no study has compared these methods to evaluate change on WAIS-IV performance. A secondary purpose of the study is to evaluate the relative accuracy of these RCI estimation methods.
METHOD

Participants

Notices of the research study were posted throughout a university campus to recruit participants, and students and individuals from the campus were solicited. Participants included 2 individuals from the community and 67 students from the university. Those from the community were compensated with a $20 gift card upon completion of both administration sessions. Participants who were students were given the option of receiving the $20 gift card or research credit for their psychology courses. The use of gift cards and research credit was approved by the local Institutional Review Board. All participants signed an informed consent form that described the purpose of the study and their rights as research participants. Prior to each assessment session participants were screened for neurological and psychiatric disease, head injury, learning disabilities, and other medical illnesses. These conditions parallel the exclusion criteria employed in recruiting the WAIS-IV normative sample (Coalson & Raiford, 2008).

Initially 93 participants agreed to participate in the study. After careful screening 24 were excluded due to the presence of at least one exclusion criterion. Among the 69 individuals who were enrolled in the study, 15 were men and 54 were women. Ages ranged from 18 to 32 with a mean age of 21.1 (SD = 2.4) at baseline. The mean education level for the group was 14.9 years (SD = 1.0). The group was comprised of 51 Caucasians, 6 African-Americans, 4 Asian-Americans, 3 Hispanics, 3 Native Americans, 1 Indian, and 1 person of multi-ethnicity. Table 1 indicates the demographic characteristics of the 69 participants enrolled in the study.

Of the 69 participants who were assessed at baseline, 54 returned for the follow-up assessment. In order to evaluate the possibility of systematic attrition or selection biases between those who did and did not complete the study, univariate analyses of variance (ANOVAs) were conducted. The groups were similar with respect to age, \( F(1, 69) = 0.74, p > .05 \), and education, \( F(1, 69) = 0.22, p > .05 \). Non-parametric statistics revealed that the group of study-completers were similar in ethnicity to those who failed to return (\( p > .05 \)). Descriptive statistics for these groups appear in Table 1. Additionally, to determine whether those who completed and failed to complete the study differed in WAIS-IV performance, a one-way

| Table 1. Demographic characteristics of participants who completed and did not complete the study |
|---------------------------------------------------------------|-----------------|-----------------|-----------------|-----------------|
| Gender                                                        | 3-month group   | 6-month group   | Completed       | Did not complete |
|                                                               | 28 Females      | 18 Females      | 46 Females      | 8 Females       |
|                                                               | 1 Male          | 7 Males         | 8 Males         | 7 Males         |
| Age                                                           | 20.9 (2.5)      | 20.9 (2.5)      | 20.9 (2.5)      | 21.5 (2.3)      |
| Education                                                     | 15.1 (1.1)      | 14.8 (0.9)      | 14.9 (1.1)      | 14.8 (1.0)      |
| Ethnicity                                                     | 1 Af-American   | 2 Af-American   | 3 Af-American   | 3 Af-American   |
|                                                               | 2 As-American   | 1 As-American   | 3 As-American   | 2 As-American   |
|                                                               | 1 Hispanic      | 2 Hispanic      | 3 Hispanic      | 1 Multi-ethnicity|
|                                                               | 1 Nat-American  | 2 Nat-American  | 3 Nat-American  |                 |

ANOVA was computed for each index score. As revealed in Table 2, no significant differences emerged.

Of the 54 who completed the two assessment intervals, participants were randomly assigned to be retested either 3 or 6 months later. Composition of the 3- and 6-month groups appears in Table 1. Non-parametric tests showed the 3-month and 6-month retesting groups did not differ according to ethnicity, $X^2(4) = 1.91$, $p = .75$. Regarding education level, those retested at 3 months were as well educated as those re-evaluated 6-months later, $F(1, 52) = .87$, $p = .35$, partial $\eta^2 = .02$. Likewise the groups failed to differ according to age, $F(1, 52) = .01$, $p = .93$, partial $\eta^2 = .00$. Nonetheless, the two groups differed in their gender, $X^2(1) = 6.41$, $p = .01$. The 3-month group included fewer men than the 6-month group. Descriptive statistics for these groups appear in Table 1.

**Materials and procedures**

At baseline and follow-up participants were administered the WAIS-IV (Wechsler, 2008). Clinical psychology doctoral students administered the WAIS-IV in accordance with the test’s standardization rules. Participants were provided the same instructions during both testing sessions, and test administrators adhered to the standardization instructions contained within the WAIS-IV manual. As part of these instructions participants were encouraged to give their best effort. All testing sessions occurred within a quiet laboratory space, in accordance with recommendations contained within the WAIS-IV administration and scoring manual. Upon completing the testing sessions participants were debriefed. They received no information concerning their performance.

These students were trained and supervised by the second author, a board-certified clinical neuropsychologist. To ensure that tests were reliably administered in accordance with WAIS-IV standardization instructions, the test administrator received training from the second author and had several years of experience administering the test. All protocols were scored twice, and inconsistencies were clarified with the second author. All data were entered twice, and errors were identified and eliminated. SPSS (version 17) was used to conduct the analyses.

| Table 2. Wechsler Adult Intelligence Scale-IV performance of participants who completed and did not complete the study |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Index score     | Completed $(n = 54)$ | Did not complete $(n = 15)$ | $F$-value | $p$-value | Partial eta squared |
| FSIQ            | 111.6 (12.1)     | 105.7 (12.6)     | 2.77     | .10     | .04             |
| VCI             | 113.6 (14.1)     | 106.2 (13.3)     | 3.37     | .07     | .05             |
| PRI             | 106.1 (13.2)     | 102.0 (12.3)     | 1.73     | .20     | .03             |
| WMI             | 103.9 (13.6)     | 101.3 (11.1)     | 0.44     | .51     | .00             |
| PSI             | 113.0 (12.6)     | 109.3 (14.3)     | 0.72     | .40     | .01             |
| GAI             | 112.0 (13.0)     | 105.0 (12.7)     | 3.50     | .07     | .10             |

FSIQ: Full Scale IQ; VCI: Verbal Comprehension Index; PRI: Perceptual Reasoning Index; WMI: Working Memory Index; PSI: Processing Speed Index; GAI: General Ability Index. Standard deviations appear in parentheses. For the $F$-values, $df = (1, 69)$. 


To assess whether the 3-month and 6-month groups displayed differential performance across time, a 2 (testing interval: 3-month vs 6-month group) \times 2 (time: baseline vs re-evaluation) \times 6 (scales: FSIQ, VCI, PRI, WMI, PSI, and General Ability Index Score; GAI) mixed-factor analysis of covariance (ANCOVA) was conducted, with time and scales being repeated factors. Average performance across these factors is depicted in Table 3. Because the groups differed according to gender composition, it was entered as a covariate. To control for Type I error, a modified Bonferroni correction was employed. Specifically, a $p$-value of .01 or less was used as the criterion for statistical significance throughout the study. Gender was not significant, $F(1, 51) = 3.77, p > .05$, partial $\eta^2 = .07$, and it was dropped as a covariate in subsequent analyses. The main effects of time, $F(1, 52) = 74.23, p < .001$, partial $\eta^2 = .59$, and scales, $F(5, 260) = 16.39, p < .001$, partial $\eta^2 = .24$, were significant. Notably, neither the main effect of testing interval nor its interactions with time or scales were significant, and the effect sizes involving this factor were small (partial $\eta^2 = .03$). Thus changes in WAIS-IV performance were not influenced by whether re-administration occurred 3 or 6 months later; score fluctuations were equivalent across these intervals.

The interaction of scales and time was significant, $F(5, 260) = 4.18, p = .001$, partial $\eta^2 = .07$, implying that performance on some scales changed more than on others. To follow up this interaction, simple main effects of time were evaluated for each scale. Table 4 indicates that participants’ performance increased on all WAIS-IV standard scores. Participants gained approximately 7, 5, 5, 4, 9, and 6 points on FSIQ, VCI, PRI, WMI, PSI, and GAI, respectively. PSI had the greatest increase among the WAIS-IV scales.

With respect to the WAIS-IV subtests, a 2 (testing interval) \times 2 (time) \times 15 (subtests) mixed-factor ANCOVA was conducted with time and subtests being repeated factors. Similar to the analysis of the IQ and factor scores, gender was also entered as a covariate due to its significant difference between the two groups. The analysis of the ANCOVA revealed that gender was not significant, $F(1, 51) = 3.05; p > .05$; partial $\eta^2 = .06$. Therefore it was excluded from further analyses. The main

### Table 3. Mean Wechsler Adult Intelligence Scale-IV scores for 3- and 6-month groups

<table>
<thead>
<tr>
<th>Index</th>
<th>Baseline</th>
<th>Retest</th>
<th>Baseline</th>
<th>Retest</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>112.4 (8.4)</td>
<td>119.1 (11.5)</td>
<td>110.5 (15.4)</td>
<td>117.9 (16.7)</td>
</tr>
<tr>
<td>VCI</td>
<td>114.2 (12.8)</td>
<td>118.4 (14.4)</td>
<td>112.8 (15.4)</td>
<td>118.6 (15.9)</td>
</tr>
<tr>
<td>PRI</td>
<td>107.4 (9.1)</td>
<td>111.0 (10.0)</td>
<td>106.4 (16.8)</td>
<td>112.4 (14.7)</td>
</tr>
<tr>
<td>WMI</td>
<td>105.7 (11.4)</td>
<td>108.9 (12.2)</td>
<td>101.6 (15.6)</td>
<td>106.0 (16.9)</td>
</tr>
<tr>
<td>PSI</td>
<td>112.9 (10.7)</td>
<td>123.4 (13.6)</td>
<td>112.0 (14.6)</td>
<td>118.9 (15.9)</td>
</tr>
<tr>
<td>GAI</td>
<td>112.0 (9.6)</td>
<td>116.5 (11.7)</td>
<td>111.1 (16.3)</td>
<td>117.9 (16.1)</td>
</tr>
</tbody>
</table>

FSIQ: Full Scale IQ; VCI: Verbal Comprehension Index; PRI: Perceptual Reasoning Index; WMI: Working Memory Index; PSI: Processing Speed Index; GAI: General Ability Index. Standard deviations appear in parentheses.
effects of time, $F(1, 52) = 95.94; \ p < .001$; partial $\eta^2 = .65$, and subtests, $F(14, 728) = 10.14; \ p < .001$; partial $\eta^2 = .16$, were significant. As with the IQ and composite scores, neither the main effect of testing interval nor its interactions with time or scales were significant, and the effect sizes involving this factor were small (partial $\eta^2 = .02$). Therefore changes in WAIS-IV subtests were not influenced by whether re-administration occurred 3 or 6 months later; score fluctuations were equivalent across these intervals.

The interaction of subtests and time was significant, $F(14, 728) = 3.54; \ p < .001$; partial $\eta^2 = .06$. To follow up this interaction, the simple main effects of time were evaluated for each subtest. Table 5 reveals that all subtest scaled scores increased during re-assessment. However, this increase was not significant for the Matrix Reasoning, Letter-Number Sequencing, Figure Weights, and Cancellation subtests.
Reliable Change Indices (RCI)

The ANOVAs revealed that the participants achieved significant gains on WAIS-IV standard and scaled scores. However, such analyses do not depict whether individual score fluctuations achieved clinical meaningfulness. To provide base-rate estimates of meaningful change, RCIs were computed for the WAIS-IV standard and subtest scaled scores.

The RCIs reveal the extent to which improvements on the WAIS-IV are due to measurement error/practice effects or clinically meaningful change by the examinees (cf. Jacobson & Traux, 1991). Furthermore, reliable change estimates are frequently used when assessing neuropsychological change in clinical samples in order to determine actual meaningful change from practice effects (cf. Chelune et al., 1993).

Three reliable change estimates were computed. One method employed the standard error of the difference (SED), and this method was originally employed to determine whether patients had benefited from psychotherapy (Jacobson & Truax, 1991). It assumes that practice effects are irrelevant. This assumption is acceptable in the assessment of change on measures of personality and attitudes. However, it is untenable when measuring cognitive abilities, because individuals are expected to perform better upon re-examination consequent to practice effects (Charter, 1996). Nonetheless, the SED method has been employed in studies of neuropsychological change (cf. Chelune et al., 1993).

To compute an RCI in this manner, the SED is multiplied by 1.96 to compute a 95% confidence interval. Test scores from re-evaluation intervals are subtracted from baseline scores. If this difference exceeds the 95% confidence interval based on the SED, then clinically meaningful change is judged to have occurred. This is characterized as the simple difference method of evaluating reliable change.

Chelune et al. (1993) relied on the SED to evaluate whether meaningful change had occurred in patients who received lobectomy for seizure disorder and another sample of patients who received no surgical intervention during the same interval. After comparing raw change scores to the base-rate estimate of change based on the SED, Chelune et al. corrected for the potential error associated with practice effects. They estimated that the patients who did not undergo lobectomy would change over time due to measurement error and practice effects rather than recovery from surgery. Thus this group’s average change across time was used to estimate the effect of practice. Accordingly Chelune et al. (1993) subtracted this average gain due to practice effects from each individual’s raw difference between baseline and retest scores. If this corrected difference exceeded the base-rate estimate of change based on the SED, the person was judged as having demonstrated clinically meaningful change. In addition to Chelune et al., these SED-based methods of evaluating reliable change have been compared by other researchers (Heaton et al., 2001; Temkin et al., 1999). Uniformly, the differences that were corrected for practice provided more accurate base-rate estimates of change than those that were not.

A more complex method of computing RCIs is based on linear regression. This method assumes that baseline scores will correlate with retest scores, and practice effects will be evident. This regression method computes base-rate estimates
of changed using the standard error of prediction SEP (Charter, 1996). The SEP is computed as follows:

$$\text{SEP} = \sqrt{\text{SPY}_2(1 - \rho_{Y1Y2})^{1/2}}$$

$\text{SD}_{Y2}$ is the standard deviation of scores during the retest interval, and $\rho_{Y1Y2}$ is the correlation between test scores across the assessment intervals. A 95% confidence interval was computed. The 95% confidence interval was computed by multiplying the SEP by $\pm 1.96$.

Estimated true scores for the retest interval were then computed. The estimated true score was computed as follows (Charter, 1996):

$$Y_{\text{true}} = M + r(Y_{\text{obs}} - M)$$

In the above equation $M$ is the sample mean of the test, $Y_{\text{obs}}$ is the actual score obtained by an individual, and $r$ is the reliability coefficient. In the current study the test–retest coefficients for each of the WAIS-IV IQ, Index, and subtest scores were used to estimate true scores upon re-evaluation. If the difference between the obtained retest score and the estimated true score exceeds the confidence intervals based on the SEP, meaningful change in performance is presumed to have occurred.

Tables 6 and 7 depict the test–retest coefficients, SEDs, SEPs, and resulting 95% confidence intervals for each IQ, Index, and subtest score for the entire sample. For the SED-based RCI method the confidence interval is the same for the corrected and uncorrected methods. The correction for practice effects is applied to the raw difference between retest and baseline scores. As mentioned previously, there were no significant differences in the rate of improvement between the 3- and 6-month re-evaluation groups, therefore reliable change index data are collapsed across these two groups.

### Table 6. Confidence intervals for Wechsler Adult Intelligence Scale-IV standard scores

<table>
<thead>
<tr>
<th>Index</th>
<th>$r_{xx}$</th>
<th>SEP</th>
<th>95% CI (SEP)</th>
<th>$\text{SEM}_{T1}$</th>
<th>$\text{SEM}_{T2}$</th>
<th>SED</th>
<th>95% CI (SED)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSIQ</td>
<td>.91</td>
<td>5.82</td>
<td>±11</td>
<td>3.63</td>
<td>4.20</td>
<td>5.55</td>
<td>±11</td>
</tr>
<tr>
<td>VCI</td>
<td>.85</td>
<td>7.88</td>
<td>±15</td>
<td>5.38</td>
<td>5.77</td>
<td>7.88</td>
<td>±16</td>
</tr>
<tr>
<td>PRI</td>
<td>.83</td>
<td>6.90</td>
<td>±14</td>
<td>5.44</td>
<td>5.11</td>
<td>7.46</td>
<td>±15</td>
</tr>
<tr>
<td>WMI</td>
<td>.92</td>
<td>5.70</td>
<td>±11</td>
<td>3.84</td>
<td>4.10</td>
<td>5.62</td>
<td>±11</td>
</tr>
<tr>
<td>PSI</td>
<td>.72</td>
<td>10.28</td>
<td>±20</td>
<td>7.24</td>
<td>7.83</td>
<td>10.66</td>
<td>±21</td>
</tr>
<tr>
<td>GAI</td>
<td>.86</td>
<td>7.06</td>
<td>±14</td>
<td>4.86</td>
<td>5.16</td>
<td>7.09</td>
<td>±14</td>
</tr>
</tbody>
</table>

FSIQ: Full Scale IQ; VCI: Verbal Comprehension Index; PRI: Perceptual Reasoning Index; WMI: Working Memory Index; PSI: Processing Speed Index; GAI: General Ability Index; CI: Confidence Interval; SEP: Standard Error of Prediction; $\text{SEM}_{T1}$: Standard Error of Measurement during the initial evaluation; $\text{SEM}_{T2}$: Standard Error of Measurement during re-evaluation; SED: Standard Error of Difference. $r_{xx}$ reflects average test–retest coefficient for the entire sample. All confidence intervals are rounded to the nearest whole digit. To use the confidence intervals, the desired confidence band should be summed with an individual’s estimated true score. Significant changes reflect the frequency of obtained scores that fell above or below the confidence interval.
Table 6 reveals that scores may fluctuate greatly across time without exceeding the 95% confidence intervals. For example, using the SEP method, retest scores may differ from the predicted true score by as much as 11, 15, and 14 points on FSIQ, VCI, and PRI. Differences between the predicted and obtained retest scores that fall within these intervals reflect random error and practice effects rather than genuine change. For the SED method, differences between the retest and baseline scores could likewise vary by as much as 11, 16, and 15 points for the FSIQ, VCI, and PRI scores, respectively, and such differences would reflect error and practice effects instead of clinically meaningful change.

Notably, using the SEP method no participant demonstrated changes on the WAIS-IV index scores or subtests that exceeded the 95% confidence interval. In contrast, using the SED-based RCI some participants demonstrated meaningful change on the WAIS-IV. With the corrected SED-based RCI method as much as 7% of the sample demonstrated gains that exceeded the 95% confidence interval on the index scores (Table 8). As much as 11% of the sample exceeded 95% confidence intervals for the subtests (Table 9). With the uncorrected SED method as much as 30% of the sample achieved gains that exceeded the 95% confidence interval on the index scores (Table 8) and 18% exceeded the 95% confidence intervals on the subtests (Table 9). Thus the SED-based RCI estimates seem prone to false positives, whereas the SEP-based estimate seems robust to such error.

To utilize the SEP-based confidence interval for an individual, the estimated true retest score is computed, and then subtracted from the actual retest score. If the difference between these scores exceeds the upper or lower limits of the 95%
confidence interval, then meaningful change is presumed to have occurred. For example, after 6 months of rehabilitation treatment a patient may obtain a FSIQ of 116, and their estimated true retest score is 104. This difference of 12 points exceeds the 11-point confidence interval (reported in Table 6), implying that clinically meaningful change has occurred.

**DISCUSSION**

Prior research reported in the WAIS-IV Technical Manual (Coalson & Raiford, 2008) examined the magnitude of practice effects over an average of
22 days. Prior to the current study no research examined the effects of practice on the WAIS-IV across a longer time interval. Consistent with expectations and prior research, the present study revealed that significant increases occurred during these intervals across all IQ and Index Scales of the WAIS-IV. The magnitude of these performance gains was considerable, and ranged from 4 to 9 points across index scores. Notably, the PSI, VCI, and PRI seemed more sensitive to the effects of practice, as score increments were considerably larger on these scales than on the WMI. Indeed, scores on the former three indices increased 9, 5, and 5 points, respectively, whereas the WMI increased only 4 points. Regarding individual subtest scaled scores, the results revealed that most scores increased during re-evaluation. However, the increases on Matrix Reasoning, Letter-Number Sequencing, Figure Weights, and Cancellation were not significant. This implies that these tests may be robust to practice effects. These findings largely parallel those reported in the WAIS-IV technical manual.

It is important to note that the participants’ increased performance is best attributed to practice effects. All participants were healthy adults. Thus changes were not due to an improved medical condition. Likewise, intelligence is presumably a stable construct, and score increments are unlikely attributable to genuine gains in intellect. To account for practice effects participants may have memorized instructions or item content during the first administration, and they recalled this information upon re-evaluation. Additionally, participants may have refined their test-taking approaches as they completed the first administration, thereby achieving better scores at follow-up (cf. Lezak, 2004). Some investigators assert that tests involving a timed component, require an infrequently practiced response, or have a single easily conceptualized solution tend to be more susceptible to practice effects. Several of these characteristics are present on the PSI subtests. For example, Symbol Search and Coding comprise the PSI, and they consist of a timed component which requires an infrequent practice response.

Subtests on the other factors scores also increased substantially, and share similar susceptibilities to practice effects. For instance, the subtests that comprise the PRI have a timed component and require a novel response from the examinee (e.g., Block Design, Visual Puzzles, and Picture Completion). Likewise subtests from the VCI may also be characterized as requiring distinctive and uncommonly performed responses (e.g., Similarities), and some items are probably susceptible to memorization (e.g., Information). Therefore, upon re-assessment, the examinee may recall these unique tasks from the previous administration that results in a reframing of their cognitive strategies in responding to test items which is indicative of practice effects.

The current study also evaluated whether magnitude of practice effect varied according to length of inter-test interval. In particular the current study evaluated whether score increments vary between 3- and 6-month inter-test intervals. No moderating effect of retest interval occurred. This parallels the findings of Basso et al. (2002) who conducted a similar study of WAIS-III performance. In that study WAIS-III gains were equivalent across 3- and 6-month retest intervals. Therefore increases in IQ and Index scores tend to remain stable across 3- and 6-month test intervals. As already mentioned, the increases reported here are essentially equivalent to those reported in the WAIS-IV technical manual, in which the
inter-test interval averaged 3 weeks. Although these findings may suggest that the magnitude of practice effects on the WAIS-IV are somewhat invariable, it remains uncertain whether the magnitude of score gains observed in this study will generalize to longer test intervals such as 12 or 24 months. Perhaps score gains would be attenuated over a lengthier retest interval because of decreasing recollection of the test over time.

The results of the current study have implications for research involving the re-administration of the WAIS-IV. Due to the susceptibility of practice effects on this testing instrument, detection of genuine change in clinical status may be obscured. McCaffrey and Westervelt (1995) recommend using dual baseline assessments in order to address this issue. They suggest that during re-administration practice effects may peak and any other changes that occur in performance are the result of treatment. Moreover, the incorporation of a no-treatment control group in treatment studies may yield indirect estimates of practice effects. The change in performance among a no-treatment group might provide an estimate of variation due to measurement error and practice effects.

In regards to clinical practice, the use of multiple baselines or no-treatment control groups are not feasible and irrelevant. However, the RCI data may provide some utility. Particularly, the estimates from the RCI may be used as a benchmark for assessing meaningful changes in performance among patients from practice effects and measurement error. However, it should be noted that the RCI data may not be optimal for clinical samples. By definition, practice effects require memory of previous testing. Many clinical patients tend to be amnesic and therefore might be less able to remember details and strategies used for the WAIS-IV than did the neurologically normal individuals tested in the current study using the WAIS-IV (see Groth-Marnat, Gallagher, Hale, & Kaplan, 2000). As a result clinical samples may have smaller gains in WAIS-IV scores across time. Furthermore, the diverse problems seen in clinical samples (e.g., traumatic brain injuries, cerebral vascular accidents) may conflate the contributions of practice effects and actual clinical change. Due to recovery and re-administration occurring at relatively the same time it may be difficult to discern practice effects from meaningful improvements. Therefore the current data may not present an ideal method for estimating reliable change in clinical samples; however it may provide a reasonable option.

Regarding the accuracy of the RCIs, the SEP-based estimates appear to be more accurate than those based on the SED. Because the sample involved a healthy control group, score fluctuations are due to practice and test error, and there is no basis for clinically meaningful change to occur. Yet the corrected and uncorrected SED-based reliable change estimates showed that as much as 11% and 30% of the sample displayed meaningful change on the WAIS-IV, whereas the SEP-based estimate showed that none of the sample changed significantly. Thus the SED-based estimates seem prone to false-positive errors, thereby indicating they have less specificity than the SEP-based reliable change estimate. It should be acknowledged, however, that we could not evaluate sensitivity to detecting change in this study. Thus, although it seems to possess satisfactory specificity, the sensitivity of the SEP-based reliable change estimate remains unknown.

It should be noted that this study had several potential limitations. The number of participants per cell is relatively small, thereby potentially limiting
statistical power. This problem notwithstanding, effect sizes for non-significant effects were small, implying that their non-significance is not entirely due to insufficient power. The generalizability of these findings may be limited due to the sample being comprised primarily of healthy, young, Caucasian adults. The participants were relatively well educated and possessed mean baseline IQ scores that were average to very superior. Therefore it is unclear if these results will generalize to other demographic groups such as elderly, below average IQ, and ethnic minority groups. Normal young adults have been shown to display greater improvements on intelligence tests compared to older adults when retested (Horton, 1992). Furthermore, practice effects between normal young adults and normal middle-age adults on the WAIS-R have been shown to occur differently among older normal adults (Ryan, Paolo, & Brundgardt, 1992). However, these results may have utility for clinicians who administered repeated intellectual assessments to neurologically impaired patients. Patients who are reported to be cognitively impaired have been shown to benefit less from practice than normal individuals (Kvale, 1987; Shatz, 1981). Therefore it is important for clinicians working with this population group to be aware that any changes in IQ functioning across 3- or 6-month intervals may be the result of genuine change as well as practice effects.

Another potential limitation to the generalizability of this study is its gender composition. Specifically, the sample was largely comprised of women. The selection of females may be attributed to the population from where this sample was drawn. The majority of the participants in this study were undergraduate psychology majors, and they tend to be primarily females. However, it seems highly probable that these results may be generalized to both genders. Gender differences on the WAIS-IV and its predecessors are negligible and of no practical consequence (Kaufman & Lichtenberger, 2006), and we have found no published research that reveals differential practice effects on the WAIS instruments between men and women.

Despite these potential limitations, the findings of this study have utility for clinicians and researchers who administered the WAIS-IV across 3- and 6-month intervals. Notably, the findings of this study revealed that the six WAIS-IV IQ and Index scales increased significantly across these two time intervals. These findings have implications for professionals working with individuals that need to be assessed across different time intervals. This may include neuropsychologists who are assessing for changes in level of functioning, rehabilitation programs assessing progress in patients, and/or individuals being re-assessed with the respect to the legal system. Ultimately, these findings may serve to explain changes in WAIS-IV retest scores across 3- and 6-month intervals.

REFERENCES


