ORIGINAL ARTICLE

Meaningful Change in Cognition in Multiple Sclerosis: Method Matters

L.A.S. Walker, P.D. Mendella, A. Stewart, M.S. Freedman, A.M. Smith

ABSTRACT: ***Objective:*** To determine if different methods of evaluating cognitive change over time yield measurably different outcomes. ***Methods:*** Twelve cognitively impaired patients with clinically definite Multiple sclerosis (10 relapsing-remitting, 2 secondary progressive) underwent neuropsychological testing (baseline, 6, 12 months). Data was analysed using: t-tests evaluating group differences on individual tests, group differences in composite scores, reliable change analyses at the level of the individual, and comparisons regarding number of tests failed at each time point. ***Results:*** Group t-tests on individual tests yielded no change. When tests were grouped according to theoretical constructs, analyses revealed change in processing speed. Reliable change estimates revealed that 16% of the sample deteriorated. When change was measured with respect to the number of domains affected at each time point, 58% of the sample deteriorated on at least one subtest. ***Conclusions:*** Methodology has a significant impact on interpretation of longitudinal data. In the same group of subjects, traditional group analyses documented no change in individual test scores or change on a single composite score. Analyses of individual results documented change from 16 to 58% of the sample. Advantages and disadvantages of each method were discussed. Findings have implications for interpretation of longitudinal studies.

RÉSUMÉ: Changements cognitifs significatifs dans la sclérose en plaques : la méthode d'évaluation importe. ***Objectif :*** Le but de l'étude était de déterminer si différentes méthodes utilisées pour l'évaluation longitudinale des changements cognitifs donnent des résultats dont les différences sont mesurables. ***Méthode :*** Douze patients ayant une atteinte cognitive, porteurs d'une sclérose en plaques (SP) cliniquement certaine (10 cas de SP rémittente récurrente, 2 de SP secondaire progressive), ont subi des tests neuropsychologiques (évaluation initiale et 6 et 12 mois plus tard). Les données ont été analysées au moyen du test t pour évaluer les différences de groupe pour chaque test, les différences de groupe des scores globaux, l'indice de Jacobson et Truax pour analyser l'évolution individuelle dans le temps et des comparaisons du nombre de tests échoués au moment de chaque évaluation. ***Résultats :*** Les tests t de groupe faits sur les tests individuels n'ont pas détecté de changements. Cependant, quand les tests ont été regroupés selon des notions théoriques, les analyses ont montré des changements dans la rapidité de traitement de l'information. Selon l'indice de Jacobson et Truax, 16% des patients de l'échantillon s'étaient détériorés. Quand le changement était mesuré par rapport au nombre de domaines atteints à chaque évaluation, 58% des patients de l'échantillon s'étaient détériorés selon au moins un sous-test. ***Conclusions :*** La méthode d'évaluation utilisée a un impact significatif sur l'interprétation de données longitudinales. Dans le même groupe de sujets, des analyses de groupe traditionnelles n'ont pas montré de changement dans les scores de tests individuels ou de changement d'un score global unique. Les analyses de résultats individuels ont montré des changements chez 16 à 58% des sujets de l'échantillon. Nous discutons des avantages et des désavantages de chaque méthode. Nos observations ont des implications pour l'interprétation des études longitudinales.

Can J Neurol Sci. 2011; 38: 282-288

Multiple sclerosis (MS) results in cognitive impairment in approximately half of affected individuals[1]. These deficits can occur very early in the disease course; sometimes even as the initial presenting symptom before a formal diagnosis can be made[2,3]. Although much individual variability exists, domains affected often include: attention and information processing speed, memory, visual perception and executive functioning[4]. Studying these deficits is important given the strong negative impact on quality of life[5].

Cognitive deficits often progress over time, although the rate of progression is slow, with changes being detectable in only a subset of individuals, and only after intervals of at least three years[6,7]. Amato et al[8] examined both cross-sectional and longitudinal studies. On the basis of their review they concluded that cognitive dysfunction is highly prevalent and that secondary-progressive subtypes may develop greater cognitive dysfunction than relapsing-remitting or primary progressive subtypes. Deterioration is slower than that of degenerative dementias but is similar in that it is unlikely to remit. Rates of progression are variable with the greatest risk of progression being current cognitive dysfunction. Cognitive deterioration proceeds in parallel with losses in neocortical volume[9].

From the Neuropsychology Service (LASW, PDM), On Track Program (PDM), Multiple Sclerosis Clinic (MSF), The Ottawa Hospital; Faculty of Medicine, Division of Neurology (LASW, MSF), School of Psychology (LASW, PDM, AMS), University of Ottawa; Neuropsychology Service (AS), Royal Ottawa Mental Health Centre; Memory Disorder Clinic (AS), Bruyère Continuing Care, Ottawa, Ontario, Canada.

Received June 15, 2010. Final Revisions Submitted September 22, 2010.

Correspondence to: Lisa Walker, The Ottawa Hospital, Psychology, 737 Parkdale Avenue, Main Floor, Room 49, Ottawa, Ontario, K1Y 1J8, Canada.

The choice of appropriate statistical methodology when measuring change in cognition over time has been a focus in many areas of study within the neuropsychological literature[10-17]. This has been addressed with regard to the cognitive effects of adjuvant chemotherapy[18], sport-related concussion[19], other types of traumatic brain injury[20], epilepsy surgery[21], and dementia[22] to name but a few. There has been less discussion of methodological issues in the MS literature. Outside of large-scale epidemiological studies, the majority of researchers choose to evaluate longitudinal change by evaluating group differences between baseline and follow-up using t-tests, analysis of variance, or correlational techniques.

In his 2003 chapter, Chelune[11] reviewed factors that must be considered when assessing change over time. The first of these is *bias* or systematic change in performance; the most common being a positive *practice effect*. In circumstances where prior exposure to a particular test causes one to perform better with a second exposure, interpretation of the findings can be difficult. Even if no change in scores is documented, this may represent an actual decrement in performance. Many neuropsychological tests are susceptible to the effects of practice in both patient and control samples. Other potential sources of bias include: participant variables (e.g. aging, ability, mood, medications), test variables (e.g. reliability, floor/ceiling effects) and duration of the test-retest interval. The second factor to consider when assessing change over time is *error*[11]. In statistical language this is called the standard *error of measurement* (SE_M; the distribution of random variations in a test score around a true score). Another form of error is regression to the mean. This is the tendency for follow-up scores to regress to the mean of the scores from the first distribution. Thus, if a score is spuriously high when first measured, upon retest the natural tendency would be for the score to be lower. The opposite is true for initial measurements that are spuriously low. Assessment tools that have low reliability are more likely to yield retest scores that regress back to the mean. This clearly results in difficulties interpreting findings.

Techniques to address these issues are embedded in traditional statistical methods used to assess group differences (e.g. t-tests for correlated groups; repeated measures ANOVA). However, when one considers the test performance of a particular individual, different statistical techniques are required. Two main statistical methods that have been utilized to measure change over time at the level of the individual are: Reliable Change Index (RCI) and Standardized Regression Based Change Scores (SRB). The two techniques vary in their ability to address the above measurement problems.

The RCI method was first introduced by Jacobson and Truax[23] when assessing change in individuals after psychotherapy treatment. The significance of the change in an individual test score is based on the difference between the baseline and retest scores for the normative subject sample. This allows the estimation of the distribution of expected change based on information found in test manuals. A change is considered unlikely to occur by chance (i.e. reliable) "if the absolute value of this change exceeds the standard deviation of the test-retest differences in the norming sample, multiplied by the z-score cut point that defines a designated percentile in the normal distribution"[16]. Because the initial RCI method was developed to evaluate change following psychotherapy, there was no consideration of practice effects, such as those encountered when a subject is retested using a cognitive measure to which they have prior exposure. Initial attempts to address this used a "constant" value for practice[14,21] but this does not consider regression to the mean. Based on suggestions by Iverson, the method was further adapted by Chelune and colleagues[24] given that measurement error may be different at baseline than at retest. Others have adapted the RCI method further by using linear regression of the retest scores on the baseline scores in the norming sample. A formula is generated that allows prediction of retest scores from any baseline score[25]. This method considers both practice and regression to the mean.

As noted, one of the problems with RCI methods is that they treat practice as a constant. However, research has demonstrated that practice effects can differ. For example, individuals with average to high average intellectual capabilities may be able to benefit more from practice than those with low average capabilities[26]. Chelune[11] also notes that the degree of change noted over time may also be impacted by the length of the test-retest interval and demographic factors such as age, education, and gender. The SRB techniques take into account measurement error, regression to the mean, differential practice effects and demographic variables. These multiple potential predictors of retest scores can be included in a multivariate regression model. Although theoretically, the regression-based techniques may be more advantageous[16], Heaton and his colleagues[12] found that classification rates of "change" or "no change" did not differ substantially between the two methods. They concluded that the simpler RCI methods that consider practice were likely more justified.

The objective of the current project was to determine if different methods of evaluating cognitive change over time yield measurably different outcomes in a sample of cognitively impaired patients with multiple sclerosis. Traditional group comparisons between baseline and retest performance were completed both for individual measures as well as composite scores grouping tests measuring similar constructs. In addition, RCI analyses (with corrections for practice) were completed to provide information on change over time in particular individuals. Finally, the method utilized by Amato and colleagues[9], which examined the number of tests failed at retest as compared to baseline, was also used.

Method

Subjects

Twelve patients with cognitive impairment, in the absence of major physical disability, were included in the analyses. Ten were diagnosed with relapsing-remitting MS and two with secondary progressive MS. Mean age was 47.3 (4.5) years (range 41-55), with a mean level of education of 15.1 (2.7) years (range 11-20). At baseline, the average number of years since diagnosis was 9.7 (6.6) (range 1-26), with an average time since last relapse of 4.2 (1.3) months (range 2-5). All patients were recruited through the MS Clinic of the Ottawa Hospital. All presented with cognitive impairment as assessed by their neurologist, themselves, or a close friend or relative.

Procedure

Informed consent form was obtained and patients were given a structured demographic interview. All completed a comprehensive neuropsychological assessment battery at baseline, 6 months and 12 months. Participants also underwent neuroimaging but these findings are beyond the scope of the current article[27].

The neuropsychological battery evaluated: premorbid intellectual functioning, processing speed, working memory, attention and concentration, executive abilities and memory. Tests administered were as follows: Quick Test[28], Wechsler Adult Intelligence Test – III (Digit Symbol, Symbol Search)[29], Wechsler Memory Scale – III (Letter-Number Sequencing, Spatial Span)[30], Gordon Diagnostic System (GDS; Vigilance and Distractibility subtests)[31], Trail Making Test[32], Auditory Consonant Trigrams[33,34], Controlled Oral Word Association Test (FAS, Animal fluency)[35], Wisconsin Card Sorting Test (WCST)[36,37], Modified Stroop Test[38], California Verbal Learning Test – II (CVLT-II short form)[39].

Data Analyses

Group differences: individual test scores

Paired-sample t-tests were performed on raw scores to evaluate group differences at baseline vs. 6 months, baseline vs. 12 months, and 6 months vs. 12 months. Corrections were made to statistically control for multiple comparisons.

Group differences: composite scores

Selected tests were grouped into four areas of cognition that are often affected in individuals with MS (see Table 1). Paired-sample t-tests were performed on composite scores (z-scores) to evaluate group differences at baseline vs. 6 months, baseline vs. 12 months, and 6 months vs. 12 months. Corrections were made to statistically control for multiple comparisons.

Individual differences: Reliable Change Index

A variation of the RCI was used that included an adjustment for practice effects that result from serial testing[18].

Table 1: Tests contributing to composite scores

Composite Domain	Tests included
Information Processing Speed	WAIS-III Processing Speed Index GDS Vigilance reaction time GDS Distractibility reaction time
Working Memory	WMS-III Working Memory Index
Susceptibility to Distraction	GDS Distractibility total correct Auditory Consonant Trigrams – 9 and 18 sec
Cognitive Flexibility	Trail Making Test – Part B WCST perseverative errors Modified Stroop IV-II

RCI = (SEdiff) (+/- 1.64) + practice effect

The *SEdiff* is the standard error of the difference which represents the spread of distribution of change scores expected had no change occurred. The *practice effect* for each variable is the mean difference between the follow-up and baseline scores. Difference scores were calculated for each subject on each measure (T2-T1, T3-T2, T3-T1). The difference scores were considered to be statistically significant and reliable at 90% confidence intervals if the degree of change fell outside the values derived from the RCI formula. Those subjects who demonstrated declines on two or more subtests were considered to have demonstrated reliable cognitive change.

Comparison of number of tests failed

Rather than focus on change in particular cognitive domains, Amato et al,[9] evaluated whether individuals with MS demonstrated deterioration in cognition overall (regardless of domain). When examining whether or not a relationship existed between cognition and neocortical volume loss, they compared the number of cognitive tests failed at each time point. A test was considered *failed* if a subject scored 2 or more standard deviations below the mean derived from a normative sample. For the purposes of the current paper, liberal criteria were used such that a test was considered failed if any subscale score fell below the mean by 2 or more standard deviations.

Results

Table 2 lists the unadjusted means and standard deviations on all measures at baseline, 6 months and 12 months. It is important to note that small practice effects were found on almost all cognitive measures.

Group differences: individual test scores

Group t-tests for scores on individual tests yielded no significant change over time.

Group differences: composite scores

When tests were grouped together according to four different theoretical constructs, analyses of resulting composite scores revealed a change in processing speed between baseline and 12 months ($t(9) = -3.81$, $p < .01$). All other comparisons were non-significant. Subjects demonstrated faster processing speed at 12 months compared to baseline.

Individual differences: Reliable Change Index

At the level of the individual, subtle cognitive losses were observed in two subjects (16% of sample). Demographic and clinical characteristics of these subjects did not differ from the overall sample. Cognitive decline was noted on measures of processing speed, memory, and executive function. Measures most helpful in identifying cognitive decline included: CVLT – II (short form; total immediate recall, short-delay free recall, long-delay free recall), phonemic fluency and semantic fluency tasks (see Figure for an example of the semantic fluency findings).

https://doi.org/10.1017/S0317167100011471 Published online by Cambridge University Press

Table 2: Mean raw scores at Baseline, 6-month and 12-month follow-up

Cognitive Domain/Measure	Baseline	6-months	12-months
Estimate of Premorbid Functioning			
Quick Test (baseline VIQ)	105.1 (11.7)		
Processing Speed			
Digit Symbol	61.8 (14.2)	64.9 (12.4)	62.1 (13.2)
Symbol Search	28.5 (8.4)	28.7 (6.7)	29.2 (6.0)
Gordon Diagnostic System (GDS)			
Vigilance reaction time	45.5 (14.3)	40.5 (6.8)	41.9 (9.4)
Distractibility reaction time	47.7 (6.5)	44.1 (4.7)	46.2 (8.2)
Trail Making – Part A	33.3 (9.4)	29.8 (7.0)	28.8 (9.9)
Trail Making – Part B	87.3 (45.7)	89.8 (59.0)	81.7 (68.9)
Working Memory			
Letter-Number Sequencing	9.4 (2.1)	9.7 (2.7)	9.3 (2.8)
Spatial Span	14.1 (3.9)	13.5 (3.8)	13.9 (4.4)
Auditory Consonant Trigrams			
9 second delay	7.5 (3.3)	7.1 (4.3)	6.3 (4.7)
18 second delay	5.2 (3.4)	4.8 (2.7)	4.8 (3.5)
Attention and Concentration			
Gordon Diagnostic System			
Vigilance commission errors	0.6 (0.8)	1.8 (4.3)	1.2 (2.2)
Vigilance total correct	28.7 (1.8)	29.6 (0.8)	28.9 (2.0)
Distractibility commission errors	5.1 (7.2)	1.5 (2.0)	3.2 (5.7)
Distractibility total correct	23.7 (5.1)	26.1 (5.1)	25.2 (5.6)
Executive Abilities			
Phonemic verbal fluency – FAS	37.3 (12.6)	36.1 (15.0)	40.8 (18.5)
Semantic verbal fluency – FAS	20.0 (6.4)	18.3 (8.5)	19.4 (8.6)
Wisconsin Card Sorting Test			
Trials administered	99.8 (16.7)	89.0 (18.4)	87.9 (22.5)
Categories completed	5.6 (1.0)	5.7 (0.9)	5.5 (1.2)
Perseverative errors	14.1 (8.0)	9.7 (9.0)	7.6 (4.3)
Perseverative responses	15.3 (8.4)	10.6 (10.5)	8.0 (4.9)
Loss of set	0.8 (1.2)	0.9 (1.2)	1.1 (1.9)
Modified Stroop Test			
I	50.8 (11.3)	50.58 (10.2)	53.9 (11.2)
II	72.7 (15.5)	70.3 (12.7)	74.8 (13.6)
III	140.8 (58.1)	134.1 (44.6)	128.6 (38.0)
IV	153.7 (48.8)	150.4 (40.4)	150.0 (34.7)
III-II	68.1 (50.9)	63.8 (36.9)	53.8 (28.8)
IV-II	81.0 (40.8)	80.1 (34.0)	75.2 (27.0)
Memory			
California Verbal Learning Test			
Trial 1	5.0 (0.7)	5.3 (1.4)	5.8 (0.9)
Trial 4	7.6 (1.2)	7.6 (1.4)	7.4 (1.0)
Total	26.9 (2.8)	26.8 (4.4)	27.8 (3.1)
Short-delay free-recall	7.0 (1.4)	6.8 (1.4)	7.5 (1.5)
Long-delay free-recall	6.1 (1.7)	5.9 (2.3)	6.8 (1.6)
Long-delay cued-recall	5.8 (2.4)	6.2 (2.4)	7.0 (1.5)
Intrusion errors	1.2 (1.5)	2.6 (2.5)	2.4 (2.1)
Repeated items	1.8 (3.0)	1.6 (1.4)	2.8 (2.3)
Recognition total correct	8.4 (0.8)	8.2 (1.2)	8.2 (1.1)
Recognition false positives	1.0 (1.0)	1.3 (1.5)	1.8 (1.6)

Comparison of number of tests failed

The methodology of Amato et al[9] revealed that 50% of subjects demonstrated a greater number of failed tests between baseline and six months, and 58% showed a greater number of failed tests between baseline and 12 months (see Table 3).

Discussion

Current data could be interpreted as documenting no change, improvement (information processing speed), or deterioration

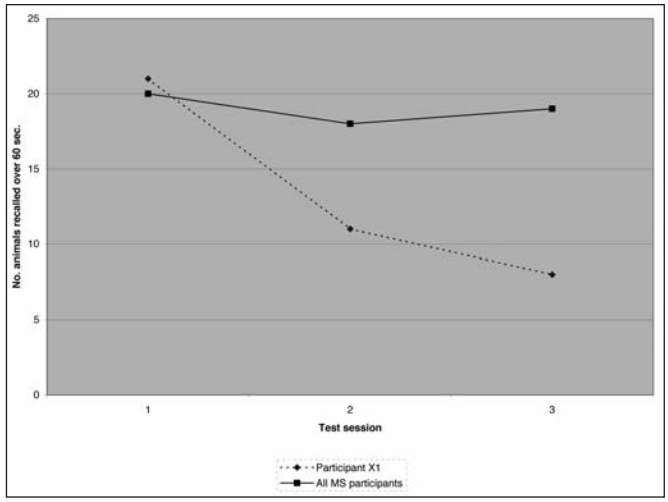

Figure: *Animal fluency performance of subject X1 compared to entire sample.*

(in a minority [16%] or a majority [58%] of subjects). This has implications for interpretation of longitudinal studies, such as when neuropsychological findings are used as outcome measures in clinical trials. There is potential for prejudice in interpretation. Each method has advantages and limitations. This paper is not meant to champion any particular method. Rather, the intention is to highlight the need to attend to methodology and consider the implications when interpreting findings.

Group analyses on individual tests are useful when attempting to measure change that occurs relatively uniformly in a sample. However, they are susceptible to increased Type I error given the multiple comparisons. In relation to MS research, the significant individual variability, which is a hallmark feature of MS, can be masked. It is difficult to identify particular sub-populations that may be vulnerable to cognitive decline when evaluated in the context of other populations that are less likely to demonstrate decline. For example, individuals with MS who present with cognitive impairment at baseline, are more likely to demonstrate cognitive decline over time when compared to those who are cognitively intact[40]. Group analysis of change may overlook such variations.

Although grouping tests according to constructs is theory-driven, individual tests included are not necessarily measuring the same construct to the same degree. In addition, most tests lack purity in that they encompass cognitive skills from multiple domains. In the current study, there was a counter-intuitive improvement in processing speed over time. Given that this group analysis did not take into account practice (comparisons to a control group were not possible), less confidence can be placed in the findings, raising questions about the value of the analysis. Although *statistically* significant, the finding is not likely clinically *meaningful*.

Analysing results at the level of the individual using reliable change methodology has the advantage of factoring out effects of practice and providing a measure of statistically meaningful change. This methodology is well suited to longitudinal research

https://doi.org/10.1017/S0317167100011471 Published online by Cambridge University Press

Table 3: Number of subjects failing a test (at least one subtest) at each timepoint

# Tests Failed	Baseline	6 months	12 months
0	3/12	3/12	1/12
1	2/12	0/12	3/12
2	1/12	3/12	2/12
3	2/12	2/12	3/12
4	3/12	2/12	1/12
5	0/12	1/12	1/12
6	1/12	0/12	1/12
7	0/12	1/12	2/12

with serial testing. It is particularly relevant to MS research given that it allows individual variability to be accounted for, while still doing so in a statistically robust manner. Although this method is well-suited to small sample sizes, it is cumbersome with larger samples, and again is susceptible to increased error associated with multiple comparisons. Iverson et al[41] highlight that RCI calculations raise the question of whether researchers should favour sensitivity or specificity. By establishing 90% confidence intervals for reliable change, this may lead to the conclusion that no change has occurred in those who have not reached this criteria. Thus, individuals may not demonstrate *statistically meaningful* change, but could very well show *clinically meaningful* change sufficient to impact quality of life. This may be particularly true for individuals who show only subtle changes that may seriously impair their ability to perform cognitively demanding jobs. Iverson et al[41] emphasized that "reliable change difference scores are meant to supplement, not replace, clinical judgment", thus real change should not be ignored clinically simply because it does not meet stringent statistical criteria.

Although the methodology utilized by Amato et al[9] was able to detect "more" change in the current sample than other methods discussed above, the specific criteria utilized may impact how clinically meaningful these changes are. Although they noted that a test was considered "failed" if a subject scored ≥ 2 standard deviations below the mean, it is unclear what "score" they used. Currently, a test was considered "failed" if any sub-score met the criteria. The subject could potentially have performed better on other aspects of the test. Base rate research shows that even healthy controls are likely to perform poorly on one or more measures within a battery of tests. Heaton et al[12] note that adequate interpretation of findings can occur only if the neuropsychologist is aware of how many changes of a given magnitude are likely to occur when a particular test battery is administered to a group of neurologically healthy individuals. This highlights the need for normative data for groups of tests when used together[42]. Specific batteries geared to MS are in widespread use (e.g. Brief Repeatable Battery of Neuropsychological Tests in Multiple Sclerosis[43]; Minimal Assessment of Cognitive Function in MS[44,45]), so this suggestion is certainly a feasible one. Using the methodology of Amato et al[9] with liberal guidelines for "failure" (i.e. failure of one subtest rather than failure on all subtests) may overestimate the prevalence of cognitive deterioration associated with MS, as it likely did in our sample. If stricter criteria were used so that tests were considered failed only if all subtest scores fell below a cut-off, then the opposite could be true (i.e.estimates of cognitive deterioration could also be underestimated). It is unclear how stringent the criteria were in the original paper[9].

Consideration of the number of tests failed at different time points provides unique information about changes in the *extent* of impairment (i.e. across cognitive domains). Ingraham and Aiken[46] stress that in a clinical context practitioners need to consider not only the number of failed tests but also the extent of the impairment and the types of tests that demonstrate impairment. Data becomes more *meaningful* when the risks and benefits of documenting change for individual patients are considered. This is particularly relevant when evaluating change as an outcome measure in clinical trials. If a negative change was falsely detected, then patients may be denied a treatment that could potentially be beneficial. Alternatively, if positive change was falsely detected patients could be offered a treatment that is essentially ineffective.

One advantage of the Amato et al[9] method over the others discussed in this paper is that individual scores are standardized and compared to normative data. Thus, whereas the first three methods assess *absolute* change over time, the fourth assesses *relative* change in comparison to a healthy population. Thus, whereas the first three methods may document change even when it does not represent a true transition into a pathological or abnormal cognitive state, the fourth can confirm whether or not true cognitive abnormality is present. Clearly then, this adds to the *meaningfulness* of the findings given that the information is more clinically relevant.

Conclusion

Although the data from the current study are interesting, one should exercise caution before drawing from it conclusions about cognition in MS in general. The small sample size certainly limits the generalizability of the findings. However, the main purpose of this paper was to generate awareness in the MS field of how important it is to consider methodology. The intent was not to promote the use of one method over another, but rather to highlight that each method has implications for

https://doi.org/10.1017/S0317167100011471 Published online by Cambridge University Press

interpretation of outcomes. Consumers of research must be vigilant to statistical methods and consider the associated advantages and limitations before generalizing conclusions.

Given the limitations of the current study with regard to both sample size and heterogeneity of subjects, future studies should attempt a similar comparison of statistical methodologies with a larger and more homogeneous sample. Application of this approach to other neurological populations may similarly raise awareness of methodological issues and the resulting implications for outcome measurement.

Acknowledgements

The authors thank the patients who kindly gave of their time to participate in this study. The research was generously funded by an unrestricted grant from Serono Canada Inc.

References

1. Rao SM, Leo GJ, Bernardin L, Unverzagt F. Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. Neurology. 1991; 41(5): 685-91.
2. Achiron A, Barak Y. Cognitive impairment in probable multiple sclerosis. J Neurol Neurosurg Psychiatry. 2003; 74: 443-6.
3. Olivares T, Nieto A, Sánchez MP, Wollmann T, Hernández MA, Barroso J. Pattern of neuropsychological impairment in the early phase of relapsing-remitting multiple sclerosis. Mult Scler. 2005; 11: 191-7.
4. Arnett PA, Higginson CI, Voss WD, Wright B, Bender WI, Wurst JM. Depressed mood in multiple sclerosis: relationship to capacity-demanding memory and attentional functioning. Neuropsychology. 1999; 13: 434-46.
5. Mitchell AJ, Benito-León J, Morales González J-M, Rivera-Navarro J. Quality of life and its assessment in multiple sclerosis: integrating physical and psychological components of wellbeing. Lancet Neurol. 2005; 4: 556-66.
6. Jennekins-Schinkel A, Laboyrie PM, Lanser JB, van der Velde EA. Cognition in patients with multiple sclerosis after four years. J Neurol Sci. 1990; 99(2-3): 229-47.
7. Kujala P, Portin R, Ruutiainen J. The progress of cognitive decline in multiple sclerosis. A controlled 3-year follow-up. Brain. 1997; 120 (Pt 2): 289-97.
8. Amato MP, Zipoli V, Portaccio E. Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies. J Neurol Sci. 2006; 245: 41-6.
9. Amato MP, Portaccio E, Goretti B, et al. Association of neocortical volume changes with cognitive deterioration in relapsing-remitting multiple sclerosis. Arch Neurol. 2007: 64(8): 1157-61.
10. Abramson IS. Reliable change formula query: a statistician's comments. J Int Neuropsychol Soc. 2000; 6: 365.
11. Chelune GJ. Assessing reliable neuropsychological change. In: Franklin RD, editor. Prediction in forensic and neuropsychology: sound statistical practices. Mahwah, NJ: Lawrence Erlbaum Associates; 2003. p. 123-47.
12. Heaton RK, Temkin N, Dikmen S, et al. Detecting change: a comparison of three neuropsychological methods using normal and clinical samples. Arch Clin Neuropsychol. 2001; 16: 75-91.
13. Hinton-Bayre A. Reliable change formula query. J Int Neuropsychol Soc. 2000; 6: 362-3.
14. Sawrie SM, Chelune GJ, Naugle RI, Lüders HO. Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. J Int Neuropsychol Soc. 1996; 2: 556-64.
15. Sawrie SM. Analysis of cognitive change: a commentary on Keith et al. Neuropsychol. 2002; 16: 429-31.
16. Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: a comparison of four models. J Int Neuropsychol Soc. 1999; 5: 357-69.
17. Temkin NR, Heaton RK, Grant I, Dikmen SS. Reliable change formula query: Temkin et al. Reply. J Int Neuropsychol Soc. 2000; 6: 364.
18. Schilling V, Jenkins V, Morris R, Deutsch G, Bloomfield D. The effects of adjuvant chemotherapy on cognition in women with breast cancer – preliminary results of an observational longitudinal study. The Breast. 2005; 14: 142-50.
19. Collie A, Maruff P, McStephen M, Darby D. Are reliable change (RC) calculations appropriate for determining the extent of cognitive change in concussed athletes? Br J Sports Med. 2003; 37: 370-6.
20. Iverson G. Interpreting change on the WAIS-III/WMS-III following traumatic brain injury. J Cog Rehab. 1999; July/August (17): 16-20.
21. Chelune GJ, Naugle RI, Lüders H, Sedlak J, Awad IA. Individual change after epilepsy surgery: practice effects and base-rate information. Neuropsychol. 1993; 7: 41-52.
22. Frerichs RJ, Tuokko HA. A comparison of methods for measuring cognitive change in older adults. Arch Clin Neuropsychol. 2005; 20: 321-33.
23. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991; 59: 12-9.
24. Chelune GJ, Sands K, Barrett J, Naugle RI, Ledbetter M, Tulsky D. Test-retest characteristics and measures of meaningful change for the Wechsler Memory Scale – III. J Int Neuropsychol Soc. 1999; 5: 109.
25. McSweeny AJ, Naugle RI, Chelune GJ, Lüders H. "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. Clin Neuropsychologist. 1993; 7: 300-12.
26. Rapport LJ, Axelrod BN, Theisen ME, Brines DB, Kalechstein AD. Relationship of IQ to verbal learning and memory: test and retest. J Clin Esp Neuropsychol. 1997; 19: 655-66.
27. Smith AM, Walker LAS, Freedman MS, DeMeuleemeester C, Hogan MJ, Cameron I. fMRI investigation of disinhibition in cognitively impaired patients with multiple sclerosis. J Neurol Sci. 2009; 281: 58-63.
28. Ammons RB, Ammons CH. The Quick Test (QT): Provisional Manual. Psychol Rep; 1962.
29. Wechsler D. Wechsler Adult Intelligence Scale-III: Administration and Scoring Manual. San Antonio, TX: The Psychological Corporation; 1997.
30. Wechsler D. Wechsler Memory Scale- III: Administration and Scoring Manual. San Antonio, TX: The Psychological Corporation; 1997.
31. Gordon M, McClure FD, Post EM. Gordon Diagnostic System. New York: Gordon Systems; 1986.
32. Army Individual Test Battery. Manual of Directions and Scoring. Washington, DC: War Department, Adjutant General's Office 1944.
33. Brown J. Some tests of the decay of immediate memory. Q J Exp Psychol. 1958; 10: 12-21.
34. Peterson LR, Peterson MJ. Short-term retention of individual verbal items. J Exp Psychol. 1959; 58: 193-8.
35. Spreen O, Benton AL. Neurosensory Center Comprehensive Examination for Aphasia (NCCEA). Victoria, BC: University of Victoria Neuropsychology Laboratory; 1969, 1977.
36. Berg EA. A simple objective technique for measuring flexibility in thinking. J Gen Psychol. 1948; 39: 15-22.
37. Grant DA, Berg EA. A behavioural analysis of degree of impairment and ease of shifting to new responses in a Weigl-type card sorting problem. J Exp Psychol. 1948; 39: 404-11.
38. Bohnen N, Jolles J, Twijnstra A. Modification of the Stroop Color Word Test improves differentiation between patients with mild head injury and matched controls. Clin Neuropsychol. 1992; 6: 1978-84.
39. Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test – Second Edition, Short Form. San Antonio, TX: The Psychological Corporation; 2000.
40. Camp SJ, Stevenson VL, Thompson AJ, et al. A longitudinal study of cognition in primary progressive multiple sclerosis. Brain. 2005; 128: 2891-8.

41. Iverson GL, Lovell MR, Collins MW. Interpreting change on ImPACT following sport concussion. Clin Neuropsychol. 2003; 17: 460-7.
42. Woods SP, Childers M, Ellis RJ, Guaman S, Grant I, Heaton RK. A battery approach for measuring neuropsychological change. Arch Clin Neuropsychol. 2006; 21: 83-9.
43. Rao SM. A manual for the Brief, Repeatable Battery of Neuropsychological Tests in Multiple Sclerosis. New York, NY: National Multiple Sclerosis Society; 1991.
44. Benedict RHB, Fischer JS, Archibald CJ, et al. Minimal neuropsychological assessment of MS patients. A consensus approach. Clin Neuropsychol. 2002; 16: 381-97.
45. Benedict RHB, Cookfair D, Gavett R, et al. Validation of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). J Int Neuropsychol Soc. 2006; 12: 549-58.
46. Ingraham LJ, Aiken CB. An empirical approach to determining criteria for abnormality in test batteries with multiple measures. Neuropsychol. 1996; 10: 120-4.

https://doi.org/10.1017/S0317167100011471 Published online by Cambridge University Press