Journal of the International Neuropsychological Society (2009), **15**, 94–102.

doi:10.1017/S1355617708090139

Impairment *versus* deficiency in neuropsychological assessment: Implications for ecological validity

NOAH D. SILVERBERG[1] AND SCOTT R. MILLIS[2]

[1]G.F. Strong Rehab Centre, Vancouver, British Columbia, Canada
[2]Department of Physical Medicine and Rehabilitation, Wayne State University School of Medicine, Detroit, Michigan

(RECEIVED March 10, 2008; FINAL REVISION September 26, 2008; ACCEPTED September 29, 2008)

Abstract

Neuropsychological test interpretation has relied on pre- and postmorbid comparisons, as exemplified by the use of demographically adjusted normative data. We argue that, when the assessment goal is to predict real-world functioning, this interpretive method should be supplemented by "absolute" scores. Such scores are derived from comparisons with the general healthy adult population (i.e., demographically unadjusted normative data) and reflect examinees' current ability, that is, the interaction between premorbid and injury/disease-related factors. In support of this view, we found that substantial discrepancies between demographically adjusted and absolute scores were common in a traumatic brain injury sample, especially in participants with certain demographic profiles. Absolute scores predicted selected measures of functional outcome better than demographically adjusted scores and also classified participants' functional status more accurately, to the extent that these scores diverged. In conclusion, the ecological validity of neuropsychological tests may be improved by the consideration of absolute scores. (*JINS*, 2009, *15*, 94–102.)

Keywords: Neuropsychological tests, Psychometrics, Activities of daily living, Work, Brain injury, Chronic, Cognition disorders

INTRODUCTION

Two core aims of neuropsychological assessment are often to determine whether a patient (a) has cognitively declined from (or returned to) their premorbid status and (b) has cognitive difficulties that are significant enough to interfere with (or sufficient to support) real-world functional task performance. These will herein be referred to as testing for *impairment* and *deficiency*, respectively. Neuropsychologists are well-equipped to address both of these questions, but typically test only for impairment, and on this basis, make (perhaps erroneous) inferences about deficiency. The main premise of this study is that detecting impairment and deficiency are distinct endeavors that require different interpretive methods.

Standard dictionaries define *impairment* as damage, reduction, or deterioration in ability, a change for the worse. It is thus relative to a baseline state. As will be familiar to the reader, impairment is determined by comparing the examinee's obtained test score with their expected premorbid score. Most commonly, their expected score is the mean of a normative group to which they are demographically similar (in a best case scenario, with respect to age, level of education attainment, gender, and ethnicity; e.g., Heaton et al., 2004). Alternatively, expected premorbid scores may be derived from examinees' current intellectual status, performance on "hold" tests that are relatively resilient to acquired brain dysfunction such as oral word reading tasks, regression equations with demographic and socioeconomic variables as predictors, historical records such as high-school grade point average, or some combination thereof (Lange et al., 2005, 2006; Miller & Rohling, 2001; Steinberg & Bieliauskas, 2005; Williams, 1997). If the standardized difference between the examinee's obtained and expected values exceeds a conventional cutoff point (e.g., < −1.5 standard deviations), the obtained score is unlikely to have been produced by this examine preinjury/predisease onset, and is, therefore, considered "impaired."[1] Of course, neuropsychologists invariably administer numerous tests, and so pre-postmorbid comparisons are made at the individual test, domain, and battery levels, cognizant of family-wise

Correspondence and reprint requests to: Noah Silverberg, G.F. Strong Rehab Centre, 4255 Laurel Street, Vancouver, British Columbia V5Z 2G9. E-mail: noah.silverberg@vch.ca

[1] In the case of abnormally distributed screening measures, standardized difference scores are inappropriate; instead, impairment is considered to be present when the obtained score is exceedingly infrequent (e.g., below the fifth percentile) in a normative sample.

Type I error inflation (Crawford et al., 2007). Pattern analysis and consideration of qualitative aspects of test performance are also essential to interpretation.

Demographic adjustments (considered as the "model" pre-postmorbid comparison method in this study) aim to statistically remove the influence of premorbid status from current test performance as to isolate the effects of injury/disease on cognition. Such normed scores, therefore, estimate the magnitude of decline from premorbid levels, irrespective of what that level was. It, therefore, seems prudent that they have been recommended for diagnostic purposes—delineating the nature and severity of cognitive difficulties owing to a known/presumed neurological condition (Heaton et al., 2004; Lezak et al., 2004; Tuokko & Woodward, 1996). They also appear appropriate to predict return to one's premorbid activities. For example, determining whether a patient can likely return to the job they held before their injury/disease onset seems most sensibly based on their degree of impairment.

In contrast to impairment, *deficiency* is defined as inadequacy, insufficiency, lack or shortage of something essential to normal functioning. Implicit is the assumption that individuals with a deficiency in a given ability cannot perform activities independently (timely and safely) that require that ability. Operationally defining deficiency, that is, "how low is too low," can be achieved by arbitrary statistical conventions (e.g., fifth percentile) or empirically, by finding the threshold that maximizes sensitivity and specificity to functional status. Importantly, whatever threshold is adopted, it must be applied to the general healthy adult population distribution and is unrelated to an individual's premorbid state. In other words, deficiency can only be determined by making "absolute" comparisons between an examinee's raw score and the grand normative mean (i.e., no demographic corrections). Absolute scores reflect the interaction between acquired brain dysfunction and premorbid status, and thus estimate current/postmorbid cognitive ability.

Authoritative sources on neuropsychological assessment (e.g., American Academy of Clinical Neuropsychology Board of Directors, 2007; Lezak et al., 2004; Mitrushina et al., 2005; Strauss et al., 2006) either make no mention or admonish the use of absolute comparisons. Some authors have argued that absolute scores are actually superior on the empirical grounds that demographically corrected scores underestimate cognitive impairment by over-adjusting for premorbid factors (e.g., Reitan & Wolfson, 2005; Yantz et al., 2006). Others have refuted this claim, concluding that demographically corrected scores "reflect more accurately the neuropsychological status of patients" (Sherrill-Pattison et al., 2000, pg. 496), are "fairer" (Steinberg & Bieliauskas, 2005, pg. 277), or are simply "better" (e.g., Taylor & Heaton, 2001, pg. 874). Capturing the sentiment of the field, Lezak et al. (2004, pg. 47) state that "most neuropsychologists assume that [correcting for all possible demographic factors] is always preferable." The central message of this article is that both types of comparisons are valuable and should both be used to address different clinical questions.

There is at least one important use for absolute comparisons: to more directly answer the question of whether a patient's abilities are sufficient for the functional demands of basic and instrumental activities of daily living (ADLs), universal employment-related tasks, specific recreational activities, etc. (cf. Heaton et al., 2004; Manly & Echemendia, 2007).[2] Each of these tasks requires a fairly uniform pattern and level of basic cognitive abilities, regardless of who performs them. In attempting to predict whether a patient is likely to have difficulty independently adhering to their medication regimen, for example, knowing their absolute level of memory ability is more important than knowing how much lower it is than a typical 55-year-old African American female with 14 years of education, and by inference, how much their memory ability has declined. As another example, whereas impairment was hypothesized to predict return to premorbid employment, deficiency should be more predictive of capacity for gainful employment of any kind, or capacity to perform specific employment-related tasks required for a new profession.

The distinction between impairment and deficiency will be trivial insofar as (1) the test variables of interest are uncorrelated with demographic factors, (2) the examinee's demographic characteristics approach the population average, or (3) the acquired brain dysfunction is extensive, thereby overwhelming the influence of demographic/premorbid factors. In such cases, neuropsychological testing will typically reveal evidence of both impairment and deficiency. However, in many common clinical situations, none of these three conditions hold, necessitating the separate consideration of impairment and deficiency. To the extent that they diverge, deficiency is purported here as conceptually more appropriate than impairment to inform patients' functional capacities. In other words, predictions of real-world functioning should be based on deficiency level. Predictions about real-world functioning and related recommendations based on impairment level may be exceedingly punitive or lax. The implication is that normative comparisons that "correct" for estimated premorbid ability (e.g., through demographic adjustments) can lower the ecological validity of neuropsychological test scores for some examinees.

This theoretical argument raises empirical questions. How discrepant are impairment (operationalized as demographically adjusted scores) and deficiency (absolute scores) levels in the same examinees? In other words, does the type of normative comparison actually make much difference? Although preferable on conceptual grounds, does deficiency actually predict functional status better than impairment? A preliminary attempt to answer these questions follows.

METHOD

Participants

Participants with traumatic brain injury were recruited through the Southeastern Michigan Traumatic Brain Injury System program. Of 62 participants, 10 were excluded from

[2] Others make a similar argument, but for age-correct scores (Mitrushina et al., 2005), e.g., "general population of 58-year-olds."

https://doi.org/10.1017/S1355617708090139 Published online by Cambridge University Press

this study because of an excess of missing data points (see below). At the time of injury, our sample of 52 participants were, on average, 38.3 years old (range, 18 to 65 years) and completed 11.5 years of education (range, 6 to 18 years); 79% were male and 73% were African-American (the remainder were Caucasian). With regard to injury severity, median Glasgow Coma Scale score at admission was 9, median loss of consciousness duration was 1.5 days, and median posttraumatic confusion duration was 24.5 days.

Materials and Procedure

All subjects were administered a neuropsychological battery for research purposes by a trained psychometrist at approximately 1 year posttraumatic brain injury. Of the tests in our research battery, five were included in the normative system of Heaton et al. (2004), and so were retained for analysis: Trail Making Test (TMT; Reitan & Wolfson, 1985), California Verbal Learning Test, Second Edition (CVLT-2; Delis et al., 2000),[3] phonemic fluency (FAS), Digit Vigilance Test (DVT; Lewis & Rennick, 1979), and Grooved Pegboard Test (GPT; Reitan & Wolfson, 1985). These tests produced seven variables: TMT Part A, TMT Part B, CVLT-2 total learning trials 1–5, CVLT-2 long delay free recall, FAS total score, DVT total time, and GPT dominant hand time (nondominant hand time was not included, as the results should be redundant).

Raw scores were converted to absolute and adjusted scores, using the tables provided in Heaton et al. (2004). Both of these are standard difference scores (expressed in T scores for the present study) derived from contrasting raw scores with normative data. Absolute scores involve a normative sample that reflects the general healthy adult population (i.e., no stratification or adjustment for any demographic variables). For the tests in our battery, this normative sample was approximately 80% Caucasian, 57% male, and averaged 49 years of age and 13.9 years of education (derived from Heaton et al., 2004); it thus reasonably corresponds to the 2000 US Census. Adjusted scores involve a normative sample that is similar to the examinee with respect to age, education, gender, and ethnicity (or technically, a predicted score from a regression model derived from healthy subjects with these demographic factors as independent variables). Overall test battery mean (OTBM) summary scores were then computed by separately averaging absolute scores and adjusted scores for the seven neuropsychological test variables. Several participants were not administered at least one test in the battery, usually due to time constraints. Those missing two or more test variables ($n=10$) were excluded from the analyses. Participants missing only one test variable ($n=6$) did not differ from those with no missing data ($n=46$) with respect to their absolute OTBM [$t(50)=.82$; $p=.42$] or adjusted OTBMs [$t(50)=-.33$; $p=.74$], and so were included (their OTBMs were computed by averaging six test variables). Discrepancy scores were computed by subtracting adjusted scores from absolute scores (i.e., positive scores indicate that the examinee's absolute score was lowered by the demographic adjustment and vice versa).

Measures of functional outcome were administered concurrently, by a rater who was blind to their neuropsychological test performance. They included the Disability Rating Scale (DRS; Rappaport et al., 1982), Glasgow Outcome Scale – Expanded (GOS-E; Wilson et al., 1998), Craig Handicap Assessment and Reporting Technique – Short Form (CHART; Whiteneck et al., 1997), and the Supervision Rating Scale (SRS; Boake, 1996). These instruments have satisfactory inter-rater reliability, sensitivity to change with recovery, and construct validity in traumatic brain injury samples (Dijkers & Greenwald, 2007; Hammon et al., 2004; Levin et al., 2001; van Baalen et al., 2006; Walker et al., 2003).

The CHART was completed with the participants. The other functional outcome measures were administered to the "best available source," the participant and/or their caregiver. Items from the measures that made no reference to the individual's premorbid ability and fell in one of the following domains were included: (1) living/residence, (2) community ambulation, (3) employment, (4) global. See Table 2 for the correspondence between the selected outcome measures and domains of functioning. DRS Employability is a seven-point ordinal scale ranging from 0 (not restricted in open labor market) to 3 (completely unemployable, even in a sheltered workshop). DRS Functioning is an 11-point ordinal scale ranging from 0 (complete independence) to 5 (totally dependent and requiring 24-hr nursing care). Responses to GOS-E items were coded as yes/no. SRS scores ranged from 1 to 13, but where dichotomized as independent living vs. receiving supervision because of severe positive skewness. CHART index scores range from a maximum of 100, "level of performance typical of the average nondisabled person," to 0, complete disability. Employment status was dichotomized as productive outcome (competitive employment) and unproductive outcome (retired/disability or unemployed/not looking); participants whose employment status was equivocal (e.g., homemaking; $n=32$) were excluded from this analysis.

This study was conducted in compliance with the Wayne State University Institutional Review Board's policy for research with de-identified databases (i.e., containing no protected health information).

RESULTS

To appreciate how much of a difference these scoring systems make, absolute minus adjusted discrepancy scores were summarized for each test and the OTBM, across subjects. As can be seen in Table 1, discrepancies greater than 1.5 standard deviations occurred in our sample. Table 1 also displays the frequency of non-trivial discrepancies in the sample, where "non-trivial" is defined as a T score difference of greater than five (0.5 standard deviations), because this moves an examinee from one category descriptor to the next (e.g.,

[3] The CVLT was actually used by Heaton et al. (2004), but the raw scores for the variables used in this study are equivalent for the CVLT and CVLT-2 (Delis et al., 2000).

https://doi.org/10.1017/S1355617708090139 Published online by Cambridge University Press

Table 1. Descriptive statistics for absolute scores, adjusted scores, and discrepancy scores

	N	Absolute Mean (*SD*)	Adjusted Mean (*SD*)	Min/Max T score difference	% with > 5 T pts difference
DVT	49	41.1 (9.7)	41.0 (10.2)	−8/+8	16.3
TMT A	52	39.7 (9.9)	39.4 (11.8)	−11/+11	28.9
TMT B	51	38.9 (11.2)	38.5 (13.0)	−13/+10	47.1
CVLT-2 Trials 1-5	52	37.7 (8.5)	34.2 (12.1)	−9/+17	57.7
CVLT-2 LDFR	52	38.1 (9.1)	36.0 (13.4)	−12/+14	44.2
FAS	51	38.8 (8.3)	40.2 (9.3)	−16/+7	23.5
GPT-dom	51	37.4 (9.8)	38.7 (10.9)	−13/+9	39.2
OTBM	52	38.7 (6.8)	38.0 (8.0)	−9/+9	36.5

Note. DVT, Digit Vigilance Test; TMT, Trail Making Test; CVLT, California Verbal Learning Test, Second Edition (CVLT-2); LDFR, long delay free recall; FAS, F-A-S version of phonemic fluency; GPT-dom, Grooved Pegboard Test, dominant hand; OTBM, overall test battery mean.

Below Average to Mildly Impaired) in the Heaton et al. (2004) system. In our sample, participants with non-trivial OTBM discrepancies were similar to those with trivial OTBM discrepancies with respect to age [$t(44) = 0.50$; $p = .62$], gender [χ^2 (1) = .73; $p = .39$], and education [$t(44) = 1.62$; $p = .11$], but more likely to be Caucasian [χ^2 (1) = 7.72; $p = .005$].

A linear regression analysis was then performed to elucidate the "risk factors" for OTBM discrepancies. Absolute-*versus*-adjusted OTBM discrepancy scores were regressed onto demographic variables: age, education, gender, and ethnicity. The overall model was significant, $F(4,41) = 34.69$; $p < .001$, adjusted $R^2 = .750$. Age (beta = −.637; $p < .001$), education (beta = .284; $p = .001$), gender (beta = −.204; $p = .01$), and ethnicity (beta = −.441; $p < .001$) all uniquely contributed to the model, with minimal colinearity (tolerance > .90 for all variables). That is, a hypothetical young highly educated Caucasian female is most likely to have a large absolute > adjusted discrepancy. In contrast, an older poorly educated African-American male is most likely to have a large absolute < adjusted discrepancy.

Next, to contrast the ability of the absolute OTBM *versus* the adjusted OTBM to predict clinical ratings of real-world functioning, we ran pairs of regression models with each OTBM as the sole predictor. The appropriate generalized linear model response family was selected for each outcome variable: binary logistic for dichotomous variables (GOS-E, employment status, SRS), ordered logistic for ordinal variables (DRS), and ordinary least squares model for continuous variables (CHART). For one of the DRS variables (Global), the proportional-odds assumption of ordered logistic regression did not hold [χ^2 (5) = 16.99; $p = .005$ for absolute OTBM], and so generalized ordered logistic regression was used instead. Due to violations of the assumptions of homoscedasticity and normally distributed residuals, robust errors were used for the ordinary least squares models. Canonical links were used for all models. To meaningfully compare the pairs of absolute *versus* adjusted OTBM models, the Bayesian Information Criterion (BIC) was computed for each. Lower values indicate better model fit. The magnitude of difference in BIC values between models is conventionally interpreted as follows (Hardin & Hilbe, 2007): between 0 and 2 ("weak"), 2 and 6 ("positive"), 6 to 10 ("strong"), and > 10 ("very strong"). As can be seen in Table 2, seven of the eight models indicated better fit for the absolute OTBM, and there was strong evidence of superiority of the absolute OTBM for three of eight (in the domains of global functioning and employment).

To better understand the clinical significance of absolute *versus* adjusted scores vis-à-vis ecological validity, the ordinal and continuous outcome variables (DRS, CHART) were dichotomized into complete independence/unrestricted (i.e., ceiling-level rating) *versus* varying levels of dependence/restrictions. The cut-off scores on the OTBMs were chosen to maximize overall accuracy in classifying functional status; they ranged from T < 33 to T < 40. Differences in classification accuracy when using the absolute *versus* adjusted OTBM as the predictor was examined in two ways: (1) for each outcome variable, across participants, and (2) for each participant, across outcome variables.

Table 3 shows the first of these analyses, the percentage of the sample that was classified accurately for each outcome variable, using the absolute or adjusted OTBM as the predictor. The absolute OTBM classified subjects at least as accurately as the adjusted OTBM across all outcome variables, and as much as 12.2% (median = 9.7%) higher. However, none of these pairwise differences reached statistical significance (all χ^2 tests were $p > .05$).

In the second of these analyses, we found that participants were classified correctly on more of the eight functional outcome measures when their absolute OTBM was the predictor compared to when their adjusted OTBM was the predictor (Wilcoxon Signed Ranks $Z = -3.13$; $p = .002$). As can be seen in Table 4, this statistic reflects that the absolute and adjusted OTBM resulted in identical overall classification accuracy in exactly half of the sample, and when they differed, participants were over five times more likely to be classified accurately on more functional outcome measures by their absolute OTBM than by their adjusted OTBM (85% *vs.* 15%). Not surprisingly, a more detailed analysis showed that ties in classification accuracy occurred disproportionately often in participants

https://doi.org/10.1017/S1355617708090139 Published online by Cambridge University Press

Table 2. Bayesian information criterion values for each model and their pairwise difference

Outcome variable				Predictor/OTBM		
Domain	*Measure*	Response family	*N*	*Absolute*	*Adjusted*	*(Difference)*
Living	Amount of supervision (SRS)	Logit	52	61.87	66.54	4.67
	Is the assistance of another person at home essential every day for some activities of daily living? (GOS-E #2)	Logit	44	56.03	60.96	4.93
Community Ambulation	Ability to move about effectively in his/her surroundings (CHART Mobility)	Gaussian*	49	442.27	443.58	1.31
	Are they able to travel locally without assistance? (GOS-E #4)	Logit	30	26.11	23.92	–2.19
Employment	Employment status	Logit	33	39.5	41.1	1.6
	Employability (DRS item H)	GO-Logit	52	189.2	196.1	6.9
Global	Overall level of daily functioning (DRS item G)	Ordered	52	203.7	210.4	6.7
	Ability to occupy time with paid or volunteer work, recreation/ leisure, household tasks, etc. (CHART Occupational)	Gaussian*	51	524.1	530.2	6.1

Note. OTBM = overall test battery mean; SRS = Supervision Rating Scale; GOS-E = Glasgow Outcome Scale, Extended; CHART = Craig Handicap Assessment and Reporting Technique; DRS = Disability Rating Scale; GO-Logit = generalized ordered logit.
* With robust errors.

with trivial differences (± 5 T points) between their absolute and adjusted OTBMs (64%) than those with non-trivial differences (26%), χ^2 (1) = 6.72; p = .010.

DISCUSSION

Normative comparisons in the psychometric assessment of cognitive abilities are clinical neuropsychology's defining feature (Ivnik, 2004). The composition of normative groups is, therefore, of fundamental importance. As the closest possible match to the examinee with respect to demographic variables is generally regarded as optimal (Lezak et al., 2004), demographically adjusted normative data have enjoyed almost exclusive use to address the gamut of neuropsychological referral questions. This study argues that "optimal" depends on the purpose of the comparison. If the clinician is interested in whether a patient has declined from their premorbid status, contrasting their obtained raw scores with their expected premorbid scores (based on age, education, gender, ethnicity, and any other variables that add to their prediction) is most appropriate. This type of comparison quantifies impairment—how much examinees' scores are lowered relative to their (estimated) preinjury/disease onset baseline. The degree of impairment is likely most predictive of the patient's success in returning to (or continuing) work or other premorbidly engaged-in functional activities with extraordinary or idiosyncratic cognitive demands. If, in contrast, the clinician is interested in determining whether the patient's cognitive abilities are sufficient for the demands of universal functional tasks (e.g., activities of daily living, driving a car, operating a cashier, etc.), comparing their raw scores with general healthy adult population norms, generating "absolute" scores, is most appropriate. This type of comparison quantifies deficiency—how low the examinees' scores are currently, reflecting an interaction between their premorbid baseline and brain injury/disease effects.

Data analyses from a sample of patients with postacute traumatic brain injury provided preliminary empirical support for this conceptual argument. The aims of these analyses were to explore (1) the frequency and magnitude of differences between absolute and adjusted scores (measuring deficiency and impairment, respectively), as well as the characteristics of examinees with significant differences between the two, and (2) their relative ability to predict clinical ratings of real-world functioning. With regard to the first aim, impairment and deficiency were found to diverge often. Approximately one third of participants in our sample obtained non-trivial discrepancies (> 0.5 standard deviations) in their OTBM, and cases of dissociation (i.e., impairment but no deficiency and vice versa) were seen. The "risk factors" for such discrepancies in the present sample are consistent with previous studies documenting the relationship between demographic variables and neuropsychological test performance: age, education, gender, and ethnicity all uniquely contributed to predicting discrepancy. The clinical relevance of this finding is that examinees with combinations of these variables in the same direction are particularly susceptible to large absolute *versus* adjusted discrepancies, and as a corollary, possible interpretive errors when the prediction of functional status is based on the latter. For example, an older adult with very low education may exhibit minimal impairment, engendering little concern about their

https://doi.org/10.1017/S1355617708090139 Published online by Cambridge University Press

Table 3. Classification accuracy of absolute and adjusted overall test battery mean scores

Outcome domain	Outcome measure	*N*	% Correct	
			Absolute	*Adjusted*
Living	Amount of supervision (SRS)	52	71.2	59.6
	Is the assistance of another person at home essential every day for some activities of daily living? (GOS-E #2)	44	77.3	75
Community Ambulation	Ability to move about effectively in his/her surroundings (CHART Mobility)	49	59.2	55.1
	Are they able to travel locally without assistance? (GOS-E #4)	30	93.3	93.3
Employment	Employment status	33	75.8	63.6
	Employability (DRS item H)	52	63.5	55.8
Global	Overall level of daily functioning (DRS item G)	52	71.2	59.6
	Ability to occupy time with paid or volunteer work, recreation/leisure, household tasks, etc. (CHART Occupational)	51	74.5	62.7

Note. OTBM = overall test battery mean; SRS = Supervision Rating Scale; GOS-E = Glasgow Outcome Scale, Extended; CHART = Craig Handicap Assessment and Reporting Technique; DRS = Disability Rating Scale.

driving safety, while the same raw scores correspond to moderate to severe deficiency, and likely, increased accident risk. Of note, future refinement in the prediction of premorbid neuropsychological test scores, such as with oral word reading performance (Schretlen et al., 2005; Testa, 2007), can be expected to improve the accuracy with which pre-postmorbid comparisons measure acquired brain dysfunction but further worsen the ecological validity of such data if they continue to be used to predict real-world functioning.

The second aim of this study was to provide a preliminary examination of the hypothesis that real-world functioning is more closely related to, and, therefore, better informed by, deficiency than impairment. In a series of regression models, the absolute OTBM was found to be a better predictor of clinical ratings of global real-world functioning than the adjusted OTBM. The strength of this evidence ranged from positive/minimal to strong across the outcome measures, according to standard interpretive criteria. Participants also tended to be classified more accurately as dependent/independent on these outcome measures by their absolute OTBM; these differences were most evident for *overall* classification accuracy, when the outcome measures were collapsed. The relatively modest superiority of absolute over adjusted OTBM scores in the whole sample is not surprising given that the two were highly correlated [$r(50) = .80$; $p < .001$]. The OTBMs were also highly similar (± 5 T points) for many participants (two-thirds of our sample), and indeed, classification accuracy differences were significantly less frequent in these participants. In conclusion, absolute scores appear to be more ecologically valid than adjusted scores for the average neuropsychological examinee with traumatic brain injury. This may be less true for examinees with highly similar absolute and adjusted scores (i.e., grossly equivalent impairment and deficiency).

Several limitations of this study are noteworthy. Because ethnic minority status and low education attainment were both

Table 4. Frequency of total correct classifications across the eight functional outcome measures for the absolute OTBM *versus* the adjusted OTBM

		Total correct classifications for adjusted OTBM								
		0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Total correct classifications for absolute OTBM	*0*	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	1	0	0	0	1	0	0
	3	0	0	4	0	0	0	0	0	0
	4	0	0	0	4	1	0	1	0	0
	5	0	0	0	0	4	1	1	0	0
	6	0	0	1	0	0	1	5	0	0
	7	0	1	1	0	1	0	1	14	1
	8	0	0	2	1	0	0	0	1	4

Note. Gray area, participants who were accurately classified on more of the functional outcome measures by their adjusted OTBM *versus* their absolute OTBM. Black area, participants who were accurately classified on more of the functional outcome measures by their absolute OTBM *versus* their adjusted OTBM. White diagonal, ties, or participants who had equivalent classification accuracy with both their OTBMs.

https://doi.org/10.1017/S1355617708090139 Published online by Cambridge University Press

over-represented in the present sample, generalizability is limited. Certainly, further research in other clinical conditions and samples with different demographic compositions is needed to adequately test our hypothesis. The frequency and magnitude of absolute *versus* adjusted score discrepancies are expected to be lower in a research sample (or clinical practice setting) that more closely reflects the American "sociocultural mean" (but comparable to another with a composition that differs from the national average in other ways, e.g., young highly-educated Caucasians at a university clinic). Consistent with the present findings, less significant discrepancies between absolute and adjusted scores should be associated with relatively unimpressive differential ecological validity. Future studies with larger samples could also separate out participants with trivial and non-trivial absolute/adjusted OTBM differences in their analyses, to better understand the distinct implications for ecological validity in examinees with these characteristics. Another limitation is that weaker reliability of certain functional outcome measures in our study may have resulted in increased error and reduced power relative to analyses involving functional outcome measures with stronger reliability. Therefore, conclusions about absolute scores better predicting some outcome domains (e.g., living independence) and not others (e.g., community ambulation) are premature. As well, only one, albeit the most common, method of pre-postmorbid comparison was used in this study. Although the present findings should hold for other methods (e.g., intelligence *versus* memory contrasts), this remains to be empirically demonstrated. Finally, future research will be necessary to examine the complementary hypothesis that adjusted scores better predict resumption of one's premorbid occupation than absolute scores.

These limitations notwithstanding, consideration of absolute scores in clinical practice may improve the ecological validity of neuropsychological assessments and the value of assessment-driven recommendations. This would complement, rather than contradict, other proposed methods of increasing ecological validity, such as developing new tests that more closely resemble real-world tasks (Burgess et al., 2006; Spooner & Pachana, 2006) and supplementing test data with observer rating scales (Chaytor et al., 2006; Sbordone & Guilmette, 1999). Best of all, it would require no additional administration time and minimal alteration of neuropsychologists' assessment practices. However, clinicians will be somewhat limited by the paucity of general healthy adult population normative data. Absolute scores are routinely provided by the expanded Halstead-Reitan battery normative system (Heaton et al., 2004), as well as a handful of test-specific manuals (e.g., Heaton et al., 1993). Approximations can be found in the descriptive statistics section of some test manuals or can be calculated by aggregating the means of stratified data. The descriptive statistics reported in the meta-analyses by Mitrushina et al. (2005) may also be useful in this regard. Considerable caution must be exercised in using these data for normative comparisons given that they were not developed for this purpose and may be quite disparate from population parameters. In report writing, the category descriptors should reflect the type of normative comparison made (e.g., "Mildly Impaired" would be inappropriate in reference to an absolute score). This may actually reduce the immense variability in assigning category descriptors among neuropsychologists (Guilmette et al., 2008) at least somewhat.

There are also foreseeable clinical implications of the impairment *versus* deficiency distinction other than with respect to ecological validity. For example, in forensic work, compensation in personal injury litigation based on cognitive losses, or the degree of acquired cognitive problems owing to an injury, would be best informed by impairment. On the other hand, compensation based on current employability or earning potential, needs for in-home professional support services, etc., would be better informed by deficiency. As another example, both the prodromal, or "mild cognitive impairment," phase and early dementia phase of neurodegenerative diseases are characterized by cognitive decline. Their distinction typically lies in the preservation *versus* deterioration in daily functioning (Gauthier et al., 2006). In this diagnostic scheme, individuals with premorbidly lower abilities will more often be diagnosed as having dementia, because less of a decline will be needed before functional impairment manifests (impairment and deficiency), whereas premorbidly higher-functioning individuals with the same magnitude of cognitive decline (i.e., comparable adjusted scores) will tend not to be diagnosed with dementia (impairment but no deficiency) until later. This can be thought of as another form of ascertainment bias (cf. Tuokko et al., 2003). Because the magnitude of decline, rather than current ability level, is likely more predictive of future decline (i.e., the presence of neurodegenerative disease), this adds to the argument against a "preserved daily functioning" criterion in the detection of prodromal dementia (e.g., Hallam et al., 2008).

There are also implications for research methodology. Research on the ecological validity of traditional neuropsychological tests has largely involved correlating them with various measures of real-world performance (see Chaytor & Schmitter-Edgecombe, 2003 for a review of this literature). Several such studies analyzed normed scores that are adjusted for demographical variables rather than raw scores (e.g., Crowe et al., 2004; LeBlanc et al., 2000; Ready et al., 2001) and still others did not specify. The present findings suggest that studies analyzing adjusted scores may underestimate the actual ecological validity of the tests being investigated.

Finally, it is important to emphasize that any kind of cognitive test score (or for that matter, any measure of bodily systems dysfunction) cannot be used in isolation to determine disability (Peterson, 2005). Rather, this endeavor depends on the interplay between cognitive abilities, emotional and neurobehavioral functioning, physical limitations (e.g., hemiplegia), caregiving resources, compensation strategies and assistive technologies, and other contextual/environmental factors. In other words, deficiency is more relevant than impairment for predicting functional capacity, but must be considered in concert with these factors, as well as non-neuropathogenic sources of poor cognitive test performance (Sbordone & Guilmette, 1999).

ACKNOWLEDGMENTS

The authors thank Robert Kotasek and Dr. Robin Hanks for their assistance in preparing the dataset. This study was supported by a grant from the US Department of Education-National Institute of Disability Research and Rehabilitation – The Traumatic Brain Injury Model Systems Project (H133A020515-04). There were no other funding sources and no perceived conflicts of interest.

REFERENCES

American Academy of Clinical Neuropsychology Board of Directors. (2007). American Academy of Clinical Neuropsychology (AACN) practice guidelines for neuropsychological assessment and consultation. *The Clinical Neuropsychologist*, *21*, 209–231.

Boake, C. (1996). Supervision Rating Scale: A measure of functional outcome from brain injury. *Archives of Physical Medicine and Rehabilitation*, *77*, 765–772.

Burgess, P.W., Alderman, N., Forbes, C., Costello, A., Coates, L.M., Dawson, D.R., Anderson, N.D., Gilbert, S.J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society; JINS*, *12*, 194–209.

Chaytor, N. & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, *13*, 181–197.

Chaytor, N., Scmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology*, *21*, 217–227.

Crawford, J.R., Garthwaite, P.H., & Gault, C.B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. *Neuropsychology*, *21*, 419–430.

Crowe, S.F., Mahony, K., & Jackson, M. (2004). Predicting competency in automated machine use in an acquired brain injury population using neuropsychological measures. *Archives of Clinical Neuropsychology*, *19*, 673–691.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (2000). *California Verbal Learning Test-Second Edition: Adult Version*. San Antonio, TX: The Psychological Corporation.

Dijkers, M. & Greenwald, G. (2007). Functional assessment in TBI rehabilitation. In N.D. Zasler, D.I. Katz, & R.D. Zafonte (Eds.), *Brain Injury Medicine: Principles and Practice*. New York: Demos.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R.C., Ritchie, K., Broich, K., Belleville, S.,Brodaty, H., Bennet, D., Chertkow, H., Cummings, J.L., de Leon, M., Feldman, H., Ganguli, M., Hampel, H., Schetlens, P., Tierney, M.C., Whitehouse, P., Windblad, B., & International Psychogeriatric Association Expert Conference on mild cognitive impairment. (2006). Mild cognitive impairment. *Lancet*, *367*, 1262–1270.

Guilmette, T.J., Hagan, L.D., & Giuliano, A.J. (2008). Assigning qualitative descriptions to test scores in neuropsychology: Forensic implications. *The Clinical Neuropsychologist*, *22*, 122–139.

Hallam, B.J., Silverberg, N.D., Lamarre, A.K., Mackenzie, I.R., & Feldman, H.H. (2008). Clinical presentation of prodromal frontotemporal dementia. *American Journal of Alzheimer's Disease and Other Dementias*, *22*, 456–467.

Hammond, F.M., Grattan, K.D., Sasser, H., Corrigan, J.D., Rosenthal, M., Bushnick, T., & Shull, W. (2004). Five years after traumatic brain injury: A study of individual outcomes and predictors of change in function. *NeuroRehabilitation*, *19*, 25–35.

Hardin, J.W. & Hilbe, J. (2007). Generalized Linear Model and Extensions. College Station, TX: State Press

Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.C., & Curtiss, G. (1993). *Wisconsin Card Sorting Test Manual: Revised and Expanded*. Lutz, FL: Psychological Assessment Resources, Inc.

Heaton, R.K., Miller, S., Taylor, M.J., & Grant, I. (2004). *Revised Comprehensive Norms for an Expanded Halstead-Reitan Battery: Demographically Adjusted Neuropsychological Norms for African American and Caucasian Adults*. Lutz, FL: Psychological Assessment Resources, Inc.

Ivnik, I.J. (2005). Normative psychology: A professional obligation? *The Clinical Neuropsychologist*, *19*, 159–161.

Lange, R.T., Chelune, G.J., & Tulsky, D.S. (2006). Development of WAIS-III General Ability Index Minus WMS-III memory discrepancy scores. *The Clinical Neuropsychologist*, *20*, 382–395.

Lange, R.T., Schoenberg, M.R., Chelune, G.J., Scott, J.G., & Adams, R.L. (2005). Developmental of the WAIS-III General Ability Index Estimate (GAI-E). *The Clinical Neuropsychologist*, *19*, 73–86.

LeBlanc, J.M., Hayden, M.E., & Paulman, R.G. (2000). A comparison of neuropsychological and situational assessment for predicting employability after closed head injury. *Journal of Head Trauma Rehabilitation, 15(4)*, 1022–1040.

Levin, H.S., Boake, C., Song, J., McCauley, S., Contant, C., Diaz-Marchan, P., Brundage, S., Goodman, H., & Kotria, K.J. (2001). Validity and sensitivity to change of the Extended Glasgow Outcome Scale in mild to moderate traumatic brain injury. *Journal of Neurotrauma*, *18*, 575–584.

Lewis, R. & Rennick, P.M. (1979). *Manual for the Repeatable Cognitive-Pereceptual-Motor Battery*. Grosse Point, MI: Axon.

Lezak, M.D., Howieson, D., & Loring, D. (2004). *Neuropsychological Assessment* (4th ed.). New York: Oxford University Press.

Manly, J.J. & Echemendia, R.J. (2007). Race-specific norms: Using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology*, *22*, 319–325.

Miller, L.S. & Rohling, M.L. (2001). A statistical interpretive method for neuropsychological test data. *Neuropsychology Review*, *11*, 143–169.

Mitrushina, M., Boone, K.B., Razani, J., & D'Elia, L.F. (2005). *Handbook of Normative Data for Neuropsychological Assessment* (2nd ed.). New York: Oxford University Press.

Peterson, D.B. (2005). International classification of functioning, disability, and health: An introduction for rehabilitation psychologists. *Rehabilitation Psychology*, *50*, 105–112.

Rappaport, M., Hall, K.M., Hopkins, K., Belleza, T., & Cope, D.N. (1982). Disability rating scale for severe head trauma: Coma to community. *Archives of Physical Medicine and Rehabilitation*, *63*, 118–123.

Ready, R.E., Stierman, L., & Paulsen, J.S. (2001). Ecological validity of neuropsychological and personality measures of executive functions. *The Clinical Neuropsychologist*, *15*, 314–323.

Reitan, R. & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation*. Tucson, AZ: Neuropsychology Press.

Reitan, R.M. & Wolfson, D. (2005). The effect of age and education transformations and neuropsychological test scores of

persons with diffuse or bilateral brain damage. *Applied Neuropsychology*, *12*, 181–189.

Sbordone, R.J. & Guilmette, T.J. (1999). Ecological validity: Prediction of everyday and vocational functioning from neuropsychological test data. In J.J. Sweet (Ed.), *Forensic Neuropsychology: Fundamentals and Practice*. Lisse: Swets and Zeitlinger.

Schretlen, D.J., Buffington, A.L., Meyer, S.M., & Pearlson, G.D. (2005). The use of word-reading to estimate "premorbid" ability in cognitive domains other than intelligence. *Journal of the International Neuropsychological Society*, *11*, 784–787.

Sherrill-Pattison, S., Donders, J., & Thompson, E. (2000). Influence of demographic variables on neuropsychological test performance after traumatic brain injury. *The Clinical Neuropsychologist*, *14*, 496–503.

Spooner, D.M. & Pachana, N.A. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*, *21*, 327–337.

Strauss, E., Spreen, O., & Elisabeth, S. (2006). *Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (3rd ed.). New York: Oxford University Press.

Steinberg, B. & Bieliauskas, L. (2005). Introduction to the special edition: IQ-based MOANS norms for multiple neuropsychological instruments. *The Clinical Neuropsychologist*, *19*, 277–279.

Taylor, M.J. & Heaton, R.K. (2001). Sensitivity and specificity of WAIS-III/WMS-III demographically corrected factor scores in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *7*, 867–874.

Testa, A. (2007). Regression-based norms: Historical development and current applications. Paper presented at the Thirty-Fifth Annual Meeting of the International Neuropsychological Society, Portland, Oregon.

Tuokko, H., Garrett, D.D., McDowell, I., Silverberg, N., & Kristjansson, B. (2003). Cognitive decline in high-functioning older adults: Reserve or ascertainment bias? *Aging and Mental Health*, *7*, 259–270.

Tuokko, H. & Woodward, T.S. (1996). Development and validation of a demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology*, *18*, 479–616.

van Baalen, B., Odding, E., van Woensel, M.P., & van Kessel, M.A. (2006). Reliability and sensitivity to change of measurement instruments used in a traumatic brain injury population. *Clinical Rehabilitation*, *20*, 686–700.

Walker, N., Mellick, D., Brooks, C.A., & Whiteneck, G.G. (2003). Measurement participation across impairment groups using the Craig Handicap Assessment Reporting Technique. *American Journal of Physical Medicine and Rehabilitation*, *82*, 936–941.

Whiteneck, G., Fougeyrolles, P., & Gerhart, K.A. (1997). Elaborating the model of disablement. In M. Fuhrer (Ed.), *Assessing Medical Rehabilitation Practices: The Promise of Outcomes Research*. Baltimore: Brooks Publishing.

Williams, J.M. (1997). The prediction of premorbid memory ability. *Archives of Clinical Neuropsychology*, *12*, 745–756.

Wilson, J.T., Pettigrew, L.E., & Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the Extended Glasgow Outcome Scale: Guidelines for their use. *Journal of Neurotrauma*, *15*, 573–585.

Yantz, C.L., Gavett, B.E., Lynch, J.K., & McCaffrey, M.J. (2006). Potential for interpretation disparities of Halstead-Reitan neuropsychological battery performances in a litigating sample. *Archives of Clinical Neuropsychology*, *21*, 809–817.