Psychol. Inj. and Law (2017) 10:96–103
DOI 10.1007/s12207-017-9282-x

Limited English Proficiency Increases Failure Rates on Performance Validity Tests with High Verbal Mediation

Laszlo A. Erdodi[1] · **Shayna Nussbaum**[1] · **Sanya Sagar**[1] · **Christopher A. Abeare**[1] · **Eben S. Schwartz**[2]

Received: 11 October 2016 / Accepted: 18 January 2017 / Published online: 31 January 2017

Abstract This study was designed to examine the effect of language proficiency and level of verbal mediation on failure rates on performance validity tests (PVTs). PVTs with high and low verbal mediation were administered to 80 healthy community-dwelling English-Arabic bilinguals. Digit Span and Animal Fluency were administered in both English and Arabic, in counterbalanced order, as part of a brief battery of neuropsychological tests. Participants with Arabic as their dominant language were 2 to 16 times more likely to fail PVTs with high verbal mediation compared to native speakers of English. When Digit Span and Animal Fluency were administered in the nondominant language, participants were 2 to 18 times more likely to fail the validity cutoffs. Language dominance explained between 6 and 31% of variance in dichotomized outcomes (pass/fail) on PVTs with high verbal mediation. None of the participants failed any PVTs with low verbal mediation. Limited language proficiency may result in dramatic increases in false positive rates on PVTs with high verbal mediation. Failure on a PVT administered in English to an examinee with a different linguistic background should be interpreted with great caution.

Keywords Cross-cultural neuropsychology · Performance validity testing · Bilingualism · Word Choice Test · Complex Ideational Material

✉ Laszlo A. Erdodi
lerdodi@gmail.com

[1] Department of Psychology, University of Windsor, 168 Chrysler Hall South, 401 Sunset Ave, Windsor, ON N9B 3P4, Canada

[2] Waukesha Memorial Hospital, Waukesha, WI, USA

Introduction

Consistently demonstrating one's highest, or at least typical, level of functioning is a basic assumption underlying neuropsychological assessment. Bigler (2015) called this assumption the "Achilles heel" of cognitive testing, which is a fitting metaphor that acknowledges a significant vulnerability in an otherwise strong design. This potential threat to the clinical utility of test data has become a source of ongoing controversy that has polarized the profession. On the one hand, there is a growing consensus that the credibility of test scores cannot be assumed but must be evaluated using objective, empirical measures (Bush et al. 2014; Heilbronner et al. 2009). On the other hand, concerns about the clinical and forensic interpretation of performance validity tests (PVTs) have been accumulating.

These concerns include the high cost of false positive errors, lack of a gold standard measure, unclear clinical interpretation of scores in the failing range (i.e., below-cutoff performance may indicate malingering, low motivation or the expression of a disease process), poorly understood relationship between PVTs and neuroanatomy/neurophysiology, and genuine, severe neurocognitive impairment as a confound (Bigler 2012, 2015; Boone 2013). Demographic variables like age (Lichtenstein et al. 2017; Lichtenstein et al. 2014) and education (Pearson 2009) have also been reported to influence base rates of failure on PVTs. In addition, Leighton et al. (2014) pointed out that the effect of the variability in testing paradigm, sensory modality, or other stimulus properties of PVTs on the probability of failure has not been studied systematically.

Native level English proficiency is another, less commonly examined assumption in neuropsychology. Most tests were developed for and normed on native speakers of English (NSE). The extent to which these tests provide a valid measure of

cognitive ability in individuals with limited English proficiency (LEP) was largely unknown until recent investigations. On purely rational grounds, LEP is expected to deflate performance primarily on tests with high verbal mediation (i.e., tasks for which being NSE is a fundamental requirement for examinees to demonstrate their true ability level on cognitive tests).

Surprisingly, some early studies were inconsistent with this hypothesis. Coffey et al. (2005) found that level of acculturation was significantly related to scores on the Wisconsin Card Sorting Test—an instrument that, at face value, appears to have low verbal mediation. Although acculturation is not synonymous with level of English proficiency, the two constructs are highly related. Their community sample of Mexican Americans performed poorly on all nine key variables examined (medium to very large effects) compared to the English norms. At the same time, no differences were found compared to the Spanish norms. Therefore, the authors concluded that the Wisconsin Card Sorting Test was not a culture-free measure.

Subsequent studies, however, found that the deleterious effect of LEP was limited to tests with high verbal mediation. Razani et al. (2007) reported a significant difference with large effect between their NSE and LEP samples on verbal IQ, but no difference on performance IQ or even full scale IQ. Likewise, the results of the study by Boone et al. (2007) were broadly consistent with the hypothesis that the level of verbal mediation drives the performance differences between NSE and LEP. Most of the significant contrasts within a battery of neuropsychological tests were observed on measures with high verbal mediation: Digit Span, Letter Fluency, and picture naming (medium to large effects). Surprisingly, on a measure of visual-constructional ability (the copy trial of the Rey Complex Figure Test), the LEP group outperformed NSE (medium effect), suggesting that the lower scores on the other tests are not due to inherent differences in global cognitive functioning.

Findings by Razani et al. (2007) provide further evidence that the level of verbal mediation accounts for a significant portion of variability in the neuropsychological profile of individuals with LEP. A large effect was observed on Digit Span between NSE and LEP, but no difference on Digit-Symbol Coding. In addition, between-group differences were more likely to emerge on the more difficult trials of the Trail Making Test, Stroop, and Auditory Consonant Trigrams. However, more recent investigations failed to replicate this with the Auditory Consonant Trigrams (Erdodi et al. 2016), raising the possibility that the effect of English proficiency on test performance varies not only across instruments, but also across samples. Overall, the evidence suggests that while LEP does not affect performance on nonverbal processing speed tasks, its deleterious effects are more likely to become apparent as the task demands or the level of verbal mediation increases.

The issue of performance validity and English proficiency is compounded in the interpretation of PVTs developed in North America and administered to individuals with LEP. Salazar et al. (2007) were among the first to examine the confluence of these two factors. They found that NSE outperformed the LEP group on the Reliable Digit Span, while the opposite was the case for the Rey Complex Figure Test effort equation (Lu et al. 2003). There was no difference between the groups on the Dot Counting Test (DCT; Boone et al. 2002) and the Rey Fifteen Item Test (Rey-15; Rey 1964), two free-standing PVTs specifically designed to evaluate the credibility of a response set.

Burton et al. (2012) compared the performance of Spanish-speaking examinees across settings (clinical, criminal, and civil forensic) and instruments. The Test of Memory Malingering and the Rey-15 at standard cutoffs were effective at differentiating the groups, while the DCT was not. These results further emphasize the complicated interaction between language proficiency, referral source, level of verbal mediation, and idiosyncratic stimulus properties of individual PVTs, consistent with earlier investigations that concluded that PVTs with low verbal mediation administered in Spanish are capable of distinguishing between credible and noncredible individuals (Vilar-Lopez et al. 2008).

The literature reviewed above converges in a number of tentative conclusions. First, LEP reliably deflates performance on tests with high verbal mediation. Second, the deleterious effects of LEP extend to some tests with low verbal mediation and tend to become more pronounced as task complexity increases. Third, the persistent negative findings on certain low verbal mediation tests when comparing NSE and LEP, in combination with the occasional superiority of the LEP groups, suggest that the observed differences are unlikely to be simply driven by lower language proficiency or overall cognitive functioning of individuals with LEP. Instead, they may reflect cultural differences in cognitive processing or approaches to testing. Finally, this pattern of findings results in a predictable increase in the base rates of failure, but this effect is limited to PVTs with high verbal mediation. Given the potentially high stakes of performance validity assessment (i.e., determining the credibility of an individual's overall presentation, and resultant assessment interpretation and clinical recommendations), further investigation of the topic seems warranted.

One of the notable limitations of existing research is the lack of direct comparison between performance in English and the participants' *other* (dominant) language. The present study was designed to address that. In addition to replicating elements of earlier investigations (a strategic mixture of PVTs with high and low verbal mediation) in a different, non-Spanish speaking bilingual sample, two of the tests were administered in both English and Arabic. Therefore, the concepts of "dominant language" and NSE could be conceptually separated and studied

independently. Based on the research literature reviewed above, we hypothesized that limited proficiency in the language of test administration will be associated with higher failure rate on PVTs with high verbal mediation. No difference was predicted on PVTs with low verbal mediation.

Method

Participants

Eighty healthy English-Arabic bilinguals were recruited for an academic research project through a research participant pool of a mid-sized Canadian university and the surrounding community. The study was approved by the institutional review board. Relevant ethical guidelines regulating research with human participants were followed throughout the project.

Mean age of the sample was 26.8 years ($SD = 16.0$). Mean level of education was 14.2 years ($SD = 1.7$). The majority of the participants were female (60%) and English-dominant (71.2%). Language dominance was determined based on a combination of self-reported relative language proficiency, language use pattern, and immigration history. Participants were asked to rate their proficiency in English and Arabic relative to each other, in percentage, so that the two ratings add up to 100, following the methodology described by Erdodi and Lajiness-O'Neill (2012).

For example, stable bilinguals would rate themselves as 50/50, indicating that they are equally proficient in both languages. Participants who immigrated to Canada as adults would rate themselves 40/60 or 30/70, indicating that they speak Arabic better than English. These individuals were classified as Arabic-dominant, and thus, as having LEP. Conversely, participants who were born in Canada, grew up as NSE, and had limited proficiency in Arabic would rate themselves as 60/40 or 70/30, and were classified as English-dominant.

Materials

Two PVTs with high verbal mediation (the Word Choice Test and Complex Ideational Material) and two with low verbal mediation (Rey-15 and Digit-Symbol Coding) were administered in English only. The Word Choice Test (Pearson 2009) is a free-standing PVT based on the forced choice recognition paradigm. The examinee is presented with 50 words, one at a time, at the rate of 3 s per word. The words are printed on a card and simultaneously read aloud by the examiner. After the learning trial, the examinee is presented with a card containing 50 word pairs (a target and a foil) and asked to identify the word that was part of the original list.

Given the high imageability and concreteness in combination with low word frequency (Davis 2014), discriminating between targets and foils during the recognition trial of the WCT is very easy. Even in clinical settings, credible patients tend to perform near the ceiling, with means ranging from 49.1 (Davis 2014) to 49.4 (Erdodi et al. 2014), which is comparable to the performance of university students in research settings (49.4; Barhon et al. 2015).

Complex Ideational Material is part of the Boston Diagnostic Aphasia Battery (Goodglass et al. 2001). It is a sentence comprehension task originally designed to aid in diagnosing and subtyping aphasia. The examiner asks a series of yes/no questions of increasing difficulty to evaluate the examinee's receptive language skills. Raw scores range from 0 to 12. The average performance in the normative sample was close to the ceiling ($M = 11.2$, $SD = 1.1$; Borod et al. 1980). Recent investigations revealed that in individuals without *bona fide* aphasia, a low score on Complex Ideational Material is a reliable indicator of invalid performance (Erdodi and Roth 2016; Erdodi et al. 2016).

The Rey-15 is one of the oldest stand-alone PVTs (Rey 1964). The examinee is presented with a card with five rows, each having three sequentially organized symbols for 10 s. The task is to reproduce as many of the original items as possible. Given the simplicity of the task, healthy controls produce near-perfect scores. Although performance is not immune to genuine neurological impairment, the Rey-15 is generally robust to the deleterious effects of brain injury (Lezak et al. 2012), making it suitable as a PVT (Boone 2013; Morse et al. 2013; O'Bryant et al. 2003). However, the low sensitivity of the Rey-15 has been repeatedly identified as a liability (Reznek 2005; Rüsseler et al. 2008).

The Digit-Symbol Coding subtest of the Wechsler Adult Intelligence Scales is a timed symbol substitution task measuring attention, visual scanning, and psychomotor processing speed. Although sensitive to the effect of diffuse neuropsychiatric deficits (Lezak et al. 2012), below certain cutoffs performance on Coding is confounded by invalid responding. Therefore, this test can function as an effective embedded validity indicator (Erdodi et al. 2016; Trueblood 1994).

Procedure

All tests were administered according to the standard procedures outlined in the technical manual by a trained research assistant who was fluent in both English and Arabic. Participants were instructed to perform to the best of their ability. However, they were not warned about the presence of PVTs, following recommendations based on previous empirical research on the negative effects of sensitizing examinees to the issue of performance validity (Boone 2007; Youngjohn et al. 1999).

Digit Span and Animal Fluency were administered in both languages, in counterbalanced order, once at the beginning and once at the end of the test battery. In addition to measuring

auditory attention, working memory, language skills, and processing speed, both tasks are well-established embedded validity indicators (Boone 2013; Sugarman and Axelrod 2015).

Data Analysis

The main descriptive statistics were failure rate (percent of the sample that scored below the validity cutoffs) and relative risk. Statistical significance was determined using t tests or χ^2. Effect size estimates were expressed in Φ^2. Relative risk ratios were computed to provide a single-number comparison of failure rates between the English- and Arabic-dominant groups.

Results

As a group, English-dominant participants were significantly younger ($M = 20.2$ years, $SD = 2.5$) than Arabic-dominant participants ($M = 42.9$ years, $SD = 22.9$): $t(78) = 7.44$, $p < .001$. The Arabic-dominant sample had significantly higher levels of education ($M = 15.0$ years, $SD = 1.5$) compared to their English-dominant counterparts ($M = 13.8$ years, $SD = 1.7$): $t(78) = 3.00$, $p < .01$. There was no significant difference in the gender ratio between the two groups (36.8 vs. 47.8% male): $\chi^2(1) = 0.82$, $p = .36$.

The English/Arabic Animal Fluency raw score ratio was significantly higher and more variable ($M = 2.62$, $SD = 1.55$) in the English-dominant sample as compared to the Arabic-dominant sample ($M = 0.90$, $SD = 0.29$): $t(78) = 5.25$, $p < .001$, $d = 1.54$ (very large effect). In other words, English-dominant participants generated, on average, 2.6 times more animal names in English than Arabic. Conversely, the output of Arabic-dominant participants on the English version was around 90% of their performance in Arabic. Likewise, there was a pronounced difference on the Boston Naming Test – Short Form between the English-dominant ($M = 11.8$, $SD = 2.2$) and Arabic-dominant ($M = 7.1$, $SD = 3.1$) subsamples: $t(78) = 7{,}69$, $p < .001$, $d = 1.75$ (very large effect). Performance on this test has been identified as a reliable indicator of language proficiency (Erdodi et al. 2016; Moreno and Kutas 2005). These findings provide empirical support for classifying participants into the two groups based on language dominance (i.e., English vs. Arabic) in addition to the self-rated language proficiency.

The Arabic-dominant sample was 16 times more likely to fail the Word Choice Test accuracy scores and two to three times more likely to fail Complex Ideational Material compared to the English-dominant sample. Since only Arabic-dominant participants failed the Word Choice Test time cutoff, a risk ratio could not be computed. All contrasts comparing the PVT failure rates as a function of language dominance (defined as English or Arabic) were statistically significant. No participant failed the Rey-15 or Digit-Symbol Coding in the entire sample (Table 1).

Participants were almost five times more likely to fail the age-corrected scaled score cutoff when the Digit Span task was administered in their nondominant language. This contrast was associated with a large effect size ($\Phi^2 = .14$). Compared to the dominant language, risk of failing Reliable Digit Span doubled during the nondominant language administration (medium effect size: $\Phi^2 = .06$). The relative risk was the highest on longest digits forward: nondominant language administration carried an almost 18-fold risk of failure (medium-large effect size: $\Phi^2 = .10$).

On Animal Fluency, when the task was administered in the nondominant language, participants were three to four times more likely to score in the failing range. All contrasts were statistically significant (Table 2). The effect of language dominance was more pronounced on demographically adjusted T-scores ($\Phi^2 = .31$) as compared to raw scores ($\Phi^2 = .22$). However, both of these effect size estimates fall in the very large range.

Discussion

The present study was designed to examine failure rate on PVTs with high and low verbal mediation in an English-Arabic bilingual sample. Consistent with previous reports (Boone et al. 2007; Razani et al. 2007) and our initial hypothesis, when PVTs with high verbal mediation were administered to participants with LEP, failure rates were two to 16 times higher as compared to NSE. As in earlier studies, no difference was observed on PVTs with low verbal mediation (Salazar et al. 2007), suggesting that the difference in relative risk for PVT failure between the LEP and NSE samples represents false positive errors.

When Digit Span and Animal Fluency were administered in the participant's nondominant language, they were two to 18 times more likely to fail validity cutoffs as compared to when these tests were administered in their dominant language. These within-individual contrasts provide a conceptual control condition that enhances the interpretation of the data, by redefining the independent variable from "English vs. Arabic" to "dominant vs. nondominant language." Hence, they examine the effect of language *proficiency* as presented with a specific cognitive task in a given language, regardless of the individual's language *dominance*. As such, English-dominant participants presented with a task in English are grouped together with Arabic-dominant participants presented with a task in Arabic and compared to English-dominant participants presented with a task in Arabic grouped together with Arabic-dominant participants presented with a task in English. These comparisons model the effect of the cognitive vulnerability stemming from limited language proficiency that

Table 1 Failure rates on performance validity tests administered in English only as a function of language dominance

Dominant language			WCT		CIM		Rey-15	$CD_{WAIS-III}$
			Accuracy	Time	Raw score	T-score	FR	ACSS
English	*n*		1	0	9	13	0	0
n = 57	%		1.8	0.0	15.8	22.8	0.0	0.0
Arabic	*n*		6	2	11	12	0	0
n = 23	%		26.1	8.7	47.8	52.2	0.0	0.0
		χ^2	12.2	5.08	8.97	6.58	–	–
		p	<.01	<.05	<.01	<.01	–	–
		RR	16.3	–	3.03	2.29	–	–

Raw score cutoffs were not listed in order to protect test security

WCT Word Choice Test (Pearson 2009), *Accuracy* number of words correctly recognized out of 50 – cutoff associated with a base rate of ≤25% in the overall clinical sample in the normative data (Erdodi et al. 2016; Pearson 2009), *Time* completion time for the recognition trial in seconds (cutoff suggested by Erdodi et al. 2016), *CIM* Complex Ideational Material subtest of the Boston Diagnostic Aphasia Examination, *Raw score* number of correct responses out of 12 (the liberal cutoff suggested by Erdodi et al. (2016) and Erdodi and Roth (2016) was used), *T-score* demographically adjusted score based on the norms by Heaton et al. (2004) – cutoff ≤29 (Erdodi et al. 2016; Erdodi and Roth 2016), *Rey-15 FR* Rey Fifteen-Item Test free recall trial (traditional cutoff suggested by Lezak (1995), $CD_{WAIS-III}$ *ACSS* age-corrected scaled scores on the Digit-Symbol Coding subtest of the Wechsler Adult Intelligence Scale – Third Edition (<6; Erdodi et al. 2016; Trueblood 1994), *RR* relative risk

manifests as poor performance during neuropsychological assessment. If the test in question is a PVT, the end result is a substantially higher risk of failing the cutoff and, therefore, being erroneously labeled as "invalid" or "noncredible."

Our results suggest that the inner logic of performance validity assessment (i.e., the task is so easy that below-threshold performance can be considered evidence of noncredible responding [Boone 2013; Larrabee 2012]) may not apply to PVTs with high verbal mediation administered to examinees with LEP, as predicted by Bigler (2015). In such cases, scores in the failing range are more likely to reflect limited proficiency in the language of test administration rather than invalid performance. Based on the existing evidence, PVTs with low verbal mediation appear to be appropriate to use in LEP populations (Razani et al. 2007; Salazar et al. 2007), although more research is clearly needed to better understand the interactions between LEP, performance validity, level of education and acculturation, task complexity, and level of verbal mediation (Boone et al. 2007; Coffey et al. 2005; Razani et al. 2007).

Table 2 Failure rates on performance validity tests administered in both English and Arabic as a function of language dominance

Language		Digit span $_{WAIS-III}$			Animal fluency	
		ACSS	RDS	LDF	Raw score	T-score
Dominant		8.7%	18.7%	1.2%	16.3%	23.8%
Nondominant		41.3%	39.9%	21.2%	61.2%	78.9%
	χ^2	11.0	4.48	8.12	17.2	24.4
	p	<.01	<.05	<.05	<.05	<.01
	Φ^2	.14	.06	.10	.22	.31
	RR	4.75	2.13	17.7	3.76	3.32

Raw score cutoffs were not listed in order to protect test security

WAIS-III Wechsler Adult Intelligence Scale – Third Edition, *ACSS* age-corrected scaled score (cutoff ≤6; Babikian et al. 2006; Hayward et al. 1987; Heinly et al. 2005; Trueblood 1994), *RDS* Reliable Digit Span (cutoff associated with a base rate of ≤25% in the overall clinical sample in the normative data; Babikian et al. 2006; Heinly et al. 2005; Pearson 2009), *LDF* longest digit span forward (cutoff suggested by Lezak et al. 2012), *Raw score* number of animals generated in 1 min (liberal cutoff suggested by Sugarman and Axelrod 2015; Hayward et al. 1987), *T-score* demographically adjusted score based on the norms by Heaton et al. (2004) – cutoff ≤33 (Sugarman and Axelrod 2015), *RR* relative risk

A potential weakness of the present design is the inherent differences in the signal detection profiles of the PVTs used. One could argue that the negative results on PVTs with low verbal mediation reflect the inability of these instruments to detect invalid response sets rather than convincing evidence of credible performance. Indeed, both the Rey-15 (Reznek 2005; Rüsseler et al. 2008) and the Digit-Symbol Coding (Erdodi et al. 2016) have been criticized for low sensitivity. A careful comparison of failure rate on these two PVTs relative to others partially substantiates these concerns.

In the Spanish speaking forensic sample of Burton and colleagues (2012), the base rate of failure on the Rey-15 was 47% (vs. 33% on the Test of Memory Malingering), indicating that the instrument is capable of detecting invalid responding. Therefore, in contrast with these studies, the 0% failure rate observed in the present sample may in fact reflect true negatives. In the study by Erdodi and Roth (2016), the failure rate on Digit-Symbol Coding was lower (21.2%) compared to some of the Digit Span-based PVTs (38.7–52.2%), but comparable to validity indicators embedded in Animal Fluency (18.5–32.8%). Likewise, the failure rate on Rey-15 (33.3%) was similar to that on Complex Ideational Material (26.5%).

Although the failure rate on Rey-15 (12.4%) and Coding (14.2%) were also some of the lowest in the larger scale study by Erdodi et al. (2016), they were broadly consistent with values observed on other, more robust PVTs such as Word Choice Test (27.0%), Digit Span (7.1–24.6%), Complex Ideational Material (13.1–19.4%), and Animal Fluency (11.7–21.6%). Furthermore, other empirical investigations found the Rey-15 useful at differentiating valid from invalid response sets in NSE (Morse et al. 2013; O'Bryant et al. 2003).

In addition, a more recent study by An et al. (2016) found that PVTs embedded in the Digit Span and verbal fluency tasks had a consistently higher failure rate than well-established, robust stand-alone PVTs in a sample of Canadian university students who volunteered to participate in academic research projects. Therefore, while Rey-15 and Digit-Symbol Coding might still underestimate the true rate of noncredible responding in the current sample, the markedly different failure rates between PVTs with high and low verbal mediation cannot be solely attributed to instrumentation artifacts.

An additional possible confound in the present study design is the difference in age and education between the two criterion groups: the Arabic-dominant sample was older and better educated than the English-dominant sample. While this likely reflects true population level differences (i.e., those with LEP likely immigrated to Canada as adults, and therefore, they tended to be older; in turn, older individuals had more time to advance their education), both of these demographic variables have been shown to influence performance on cognitive tests (Heaton et al. 2004; Mitrushina et al. 1999). However, our findings indicate that there was no systematic, clinically relevant difference in failure rates as a function of demographically corrected (Digit Span age-corrected scaled score, Animal Fluency T-score) vs. raw score (Reliable Digit Span, Longest Digits Forward, number of animal names generated during Animal Fluency) based cutoffs (Table 2).

There was considerable fluctuation in effect sizes associated with language dominance across instruments and cutoffs. Within Digit Span, language proficiency had the strongest relationship with the age-corrected scaled score (large effect). Conversely, the Reliable Digit Span was the least affected by language dominance (medium effect). Even though participants who were administered the test in their nondominant language were twice as likely to fail this PVT compared to when it was administered in their dominant language, the Reliable Digit Span was nevertheless the most robust validity indicator with high verbal mediation of the ones examined in the present study. However, the effect of language dominance on the likelihood of failing the Animal Fluency cutoffs was very large. As such, the use of these cutoffs in examinees with LEP is difficult to justify.

In the context of the landmark paper by Green et al. (2001) demonstrating that noncredible responding explained between 49 and 54% of variance in neuropsychological test scores, our data indicate that language proficiency accounts for 6–31% in failure rate on PVTs with high verbal mediation. Given that the base rate of failure on PVTs with low verbal mediation was zero regardless of language proficiency, most of the failures on PVTs with high verbal mediation are likely false positive errors. The implication of this finding to clinical practice is that PVTs with high verbal mediation are unreliable indicators of noncredible performance in examinees with LEP. At the same time, our data support the continued use of PVTs with low verbal mediation, provided that the examinee was able to comprehend the test instructions.

Results should be interpreted in the context of the study's limitations. As discussed above, future studies would benefit from using multiple, more sensitive PVTs, especially in the low verbal mediation category. In addition, the sample is restricted to a single geographic area and English-Arabic bilinguals. Replications based on participants with diverse ethnic and linguistic backgrounds, and using different instruments are crucial to establish the generalizability of our findings.

Despite the common variability in results across studies (Leighton et al. 2014), the cumulative evidence converges in one main conclusion: Cultural influences on neuropsychological testing are significant, vary across measures, and can significantly alter the clinical interpretation of the data. Depending on the context, the language in which the material is presented can have subtle (Erdodi and Lajiness-O'Neill 2012) or unexpectedly strong (Coffey et al. 2005) effects. Our finding that LEP can dramatically increase the failure rate on PVTs with high verbal mediation has far-reaching clinical and forensic implications, substantiates Bigler's (2012, 2015) concerns about inflated false positive rates in vulnerable populations, and warrants further investigation. In the meantime, given the high cost of misclassifying an individual as noncredible in both clinical and forensic assessments, the use of PVTs with high verbal mediation in individuals with LEP should either be avoided altogether, or interpreted with caution.

Compliance with Ethical Standards

Conflict of Interest This project received no financial support from outside funding agencies. The authors have no disclosures to make that could be interpreted as conflict of interests.

Human and Animal Rights and Informed Consent Relevant ethical guidelines regulating research involving human participants were followed throughout the project. All data collection, storage, and processing was done in compliance with the Helsinki Declaration.

References

An, K. Y., Kaploun, K., Erdodi, L. A., & Abeare, C. A. (2016). Performance validity in undergraduate research participants: a comparison of failure rates across tests and cutoffs. *The Clinical Neuropsychologist*. doi:10.1080/13854046.2016.1217046. Advance online publication.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various Digit Span scores in the detection of suspect effort. *The Clinical Neuropsychologist, 20*, 145–159.

Barhon, L. I., Batchelor, J., Meares, S., Chekaluk, E., & Shores, E. A. (2015). A comparison of the degree of effort involved in the TOMM and the ACS Word Choice Test using a dual-task paradigm. *Applied Neuropsychology: Adult, 22*, 114–123.

Bigler, E. D. (2015). Neuroimaging as a biomarker in symptom validity and performance validity testing. *Brain Imaging and Behavior, 9*, 421–444.

Bigler, E. D. (2012). Symptom validity testing, effort and neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 632–642.

Boone, K. B. (2007). *Assessment of feigned cognitive impairment. A neuropsychological perspective*. New York, NY: Guilford.

Boone, K. B. (2013). *Clinical practice of forensic neuropsychology*. New York, NY: Guilford.

Boone, K., Lu, P., & Herzberg, D. (2002). *The dot counting test*. Los Angeles: Western Psychological Services.

Boone, K. B., Victor, T. L., Wen, J., Razani, J., & Ponton, P. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology, 22*, 355–365.

Borod, J. C., Goodglass, H., & Kaplan, E. (1980). Normative data on the Boston diagnostic aphasia examination, parietal lobe battery, and the Boston naming test. *Journal of Clinical Neuropsychology, 2*, 209–215.

Burton, V., Vilar-Lopez, R., & Puente, A. E. (2012). Measuring effort in neuropsychological evaluations of forensic cases of Spanish speakers. *Archives of Clinical Neuropsychology, 27*(3), 262–267.

Bush, S. S., Heilbronner, R. L., & Ruff, R. M. (2014). Psychological assessment of symptom and performance validity, response bias, and malingering: official position of the association of psychological advancement in psychological injury and law. *Psychological Injury and Law, 7*, 197–205.

Coffey, D., Marmol, L., Schock, L., & Adams, W. (2005). The influence of acculturation on the Wisconsin card sorting test by Mexican Americans. *Archives of Clinical Neuropsychology, 20*, 795–803.

Davis, J. J. (2014). Further consideration of advanced clinical solutions word choice: comparison to the recognition memory test—words and classification accuracy on a clinical sample. *The Clinical Neuropsychologist, 28*(8), 1278–1294.

Erdodi, L. A., Abeare, C. A., Lichtenstein, J. D., Tyson, B. T., Kucharski, B., Zuccato, B. G., & Roth, R. M. (2016). Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) processing speed scores as measures of non-credible responding: The third generation of embedded performance validity indicators. Advance online publication. *Psychological Assessment*. doi:10.1037/pas0000319

Erdodi, L. A., Jongsma, K. A., & Issa, M. (2016). The 15-item version of the Boston Naming test as an index of English proficiency. *The Clinical Neuropsychologist*. doi:10.1080/13854046.2016.1224392. Advance online publication.

Erdodi, L. A., Kirsch, N. L., Lajiness-O'Neill, R., Vingilis, E., & Medoff, B. (2014). Comparing the recognition memory test and the word choice test in a mixed clinical sample: Are they equivalent? *Psychological Injury and Law, 7*(3), 255–263.

Erdodi, L., & Lajiness-O'Neill, R. (2012). Humor perception in bilinguals: Is language more than a code? *International Journal of Humor Research, 25*(4), 459–468. doi:10.1515/humor-2012-0024

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Lichtenstein, J. D., Pelletier, C. L., Rai, J. K., & Roth, R. M. (2016). The BDAE Complex Ideational Material – A measure of receptive language or performance validity? *Psychological Injury and Law, 9*, 112–120.

Erdodi, L. A., Tyson, B. T., Shahein, A., Lichtenstein, J. D., Abeare, C. A., Pelletiere, C. L., … Roth, R. M. (2016). The power of timing: Adding a time-to-completion cutoff to the Word Choice Test and Recognition Memory Test improves classification accuracy. *Journal of Clinical and Experimental Neuropsychology*. Advance online publication. doi:10.1080/13803395.2016.1230181

Erdodi, L. A., & Roth, R. M. (2016). Low scores on BDAE Complex Ideational Material are associated with invalid performance in adults without aphasia. Advance online publication. *Applied Neuropsychology: Adult*. doi:10.1080/23279095.2016.1154856

Green, P., Rohling, M. L., Lees-Haley, P. R., & Allen, L. M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury, 15*(12), 1045–1060.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston Diagnostic Aphasia Examination* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Hayward, L., Hall, W., Hunt, M., & Zubrick, S. R. (1987). Can localized brain impairment be simulated on neuropsychological test profiles? *Australian and New Zealand Journal of Psychiatry, 21*, 87–93.

Heaton, R. K., Miller, S. W., 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.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American academy of neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*(7), 1093–1129.

Heinly, M. T., Greve, K. W., Bianchini, K., Love, J. M., & Brennan, A. (2005). WAIS Digit-Span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment, 12*(4), 429–444.

Larrabee, G. J. (2012). *Forensic neuropsychology: A scientific approach*. New York: Oxford University Press.

Leighton, A., Weinborn, M., & Maybery, M. (2014). Bridging the gap between neurocognitive processing theory and performance validity assessment among the cognitively impaired: A review and methodological approach. *Journal of the International Neuropsychological Society, 20*, 873–886.

Lezak, M. D. (1995). *Neuropsychological assessment*. New York: Oxford University Press.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment*. New York: Oxford University Press.

Lichtenstein, J. D., Erdodi, L. A., Rai, J. K., Mazur-Mosiewicz, A., & Flaro, L. (2017). Wisconsin Card Sorting Test embedded validity indicators developed for adults can be extended to children. *Child Neuropsychology*. doi:10.1080/09297049.2016.1259402. Advance online publication.

Lichtenstein, J. D., Moser, R. S., & Schatz, P. (2014). Age and test setting affect the prevalence of invalid baseline scores on neurocognitive tests. *American Journal of Sports Medicine, 42*(2), 479–484.

Lu, P. H., Boone, K. B., Cozolino, L., & Mitchell, C. (2003). Effectiveness of the Rey-Osterrieth complex figure test and the Meyers and Meyers recognition trial in the detection of suspect effort. *The Clinical Neuropsychologist, 17*(3), 426–440.

Mitrushina, M. N., Boone, K. B., & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment*. New York, NY: Oxford University Press.

Moreno, E. M., & Kutas, M. (2005). Processing semantic anomalies in two languages: An electrophysiological exploration in both

languages of Spanish-English bilinguals. *Cognitive Brain Research, 22*, 205–220.

Morse, C. L., Douglas-Newman, K., Mandel, S., & Swirsky-Sacchetti, T. (2013). Utility of the Rey-15 recognition trial to detect invalid performance in a forensic neuropsychological sample. *The Clinical Neuropsychologist, 27*(8), 1395–1407.

O'Bryant, S. E., Hilsabeck, R. C., Fisher, J. M., & McCaffrey, R. J. (2003). Utility of the Trail Making Test in the assessment of malingering in a sample of mild traumatic brain injury litigants. *The Clinical Neuropsychologist, 17*(1), 69–74.

Pearson (2009). *Advanced Clinical Solutions for the WAIS-IV and WMS-IV – Technical Manual*. San Antonio, TX: Author

Razani, J., Murcia, G., Tabares, J., & Wong, J. (2007). The effects of culture on the WASI test performance in ethnically diverse individuals. *The Clinical Neuropsychologist, 21*, 776–788.

Razani, J., Burciaga, J., Madore, M., & Wong, J. (2007). Effects of acculturation on tests of attention and information processing in an ethnically diverse group. *Archives of Clinical Neuropsychology, 22*, 333–341.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Reznek, L. (2005). The Rey 15-item memory test for malingering: A meta-analysis. *Brain Injury, 19*(7), 539–543.

Rüsseler, J., Brett, A., Klaue, U., Sailer, M., & Münte, T. F. (2008). The effect of coaching on the simulated malingering of memory impairment. *BMC Neurology, 8*(37), 1–14.

Salazar, X. F., Lu, P. H., Wen, J., & Boone, K. B. (2007). The use of effort tests in ethnic minorities and in non-English speaking and English as a second language populations. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective* (pp. 405–427). New York: Guilford.

Sugarman, M. A., & Axelrod, B. N. (2015). Embedded measures of performance validity using verbal fluency tests in a clinical sample. *Applied Neuropsychology: Adult, 22*(2), 141–146.

Trueblood, W. (1994). Qualitative and quantitative characteristics of malingered and other invalid WAIS-R and clinical memory data. *Journal of Clinical and Experimental Neuropsychology, 14*(4), 697–607.

Vilar-Lopez, R., Gomez-Rio, M., Caracuel-Romero, A., Llamas-Elvira, J., & Perez-Garcia, M. (2008). Use of specific malingering measures in a Spanish sample. *Journal of Clinical and Experimental Neuropsychology, 30*(6), 710–722.

Youngjohn, J. R., Lees-Haley, P. R., & Binder, L. M. (1999). Comment: Warning malingerers produces more sophisticated malingering. *Archives of Clinical Neuropsychology, 14*, 511–515.