Personnel Assessment and Decisions

Volume 8 | Issue 1 Article 3

2022

The Effect of English Language Proficiency and Glossary Provision on Personality Measurement

Damian Canagasuriam
Saint Mary's University - Canada

Sharmili Jong
Department of National Defence

Wendy Darr
Department of National Defence

Recommended Citation
Canagasuriam, Damian; Jong, Sharmili; and Darr, Wendy (2022) "The Effect of English Language Proficiency and Glossary Provision on Personality Measurement," *Personnel Assessment and Decisions*: Number 8 : Iss. 1 , Article 3.
DOI: https://doi.org/10.25035/pad.2022.01.003
Available at: https://scholarworks.bgsu.edu/pad/vol8/iss1/3

The Effect of English Language Proficiency and Glossary Provision on Personality Measurement

Damian Canagasuriam[1], Sharmili Jong[2], and Wendy Darr[2]

1. Saint Mary's University - Canada
2. Department of National Defence - Canada

ABSTRACT

Research on English language learners suggests that language proficiency can affect the validity of standardized test scores. This study examined whether the provision of a glossary as a test accommodation during personality test completion influences the measurement of personality. Using an experimental research design, participants recruited from Amazon Mechanical Turk and Prime Panels (*n* = 206) were first categorized as having limited or high English language proficiency and then randomly assigned to a glossary condition. The results indicate that providing a within-text glossary does not impact the construct validity and reliability of personality measures. The results also suggest that participants who received glossaries found them useful. However, those who were not provided with one disagreed that they would benefit from the provision of a glossary.

KEYWORDS

personality, accommodation, glossary, language proficiency

Personality is considered to be a valuable predictor of job performance (Barrick & Mount, 1991; Barrick et al., 2001). Consequently, selection processes in some organizations incorporate assessments of personality. In such contexts, it is possible that personality measures administered in English are completed by respondents who are not proficient in the language. For example, Although English is the only language spoken by a majority of the United States population aged 18 and above, 18% reported speaking another language, and of these, 26% indicated that they spoke English "not well" or "not at all" (U.S. Census Bureau, 2000). Given evidence suggesting that language proficiency can negatively affect the construct validity of standardized test scores (Abedi et al., 2001, 2003; Cocking & Chipman, 1988), personality test scores can be similarly affected, which in turn can influence selection/hiring decisions.

Findings from the standardized achievement testing literature indicate that despite having similar test content knowledge, those with limited English proficiency (e.g., non-native speakers) performed worse on English academic tests than those with high English proficiency (e.g., native speakers; Abedi et al., 1997, 2003; Cocking & Chipman, 1988). This research suggests that at least some of the differences in standardized test scores may be explained by a reduced ability to understand test questions. This notion is supported by Abedi et al. (2003), who found that the difference between native and non-native English-speaking students' math and science standardized test scores widened as the linguistic complexity of the tests increased.

The language capability differences between native and non-native English speakers may explain differences in performance on English standardized tests. For example, non-native speakers have a smaller English vocabulary than native speakers (Umbel et al., 1992; Verhallen & Schoonen, 1993). Non-native speakers also tend to have difficulty understanding English words that are polysemous (words with more than one meaning), false cognates (root words with different meanings across languages), and abstract (words without concrete definitions; Cohen et al., 2017). Thus, standardized test scores of non-native speakers may

Corresponding author:
Wendy Darr
Author Email: wendy.darr@gmail.com

not accurately reflect their ability in the assessed domain because they may not understand certain test words and phrases (Butler & Castellon-Wellington, 2005; Martiniello, 2008).

To our knowledge, the effects of language proficiency on the construct validity of personality tests scores have never been examined. This issue is especially relevant to personality tests that are typically based on the lexical approach to construct measurement, the premise of which is that prominent individual differences can be captured in language that becomes part of everyday terminology (i.e., not needing to be defined; John et al., 1988). However, one of the limitations of this approach is that the terms may not translate similarly across languages, which can affect how individuals with limited proficiency in the test language answer test items. Thus, English personality tests completed by individuals who are not proficient in the English language may yield assessments that are less valid and/or reliable compared to those completed in one's native language. One potential solution to reduce the interference of language in personality testing is to provide a glossary of definitions to individuals who are not proficient in the test language.

Glossary Provision as an Accommodation

Test accommodations are minor modifications to standardized testing procedures meant to reduce the impact of construct-irrelevant factors (e.g., English proficiency), so that the underlying areas of interest (e.g., personality) are more accurately assessed (AERA, APA, & NCME, 2014). One type of accommodation that may help address language-related difficulties is glossaries. Unlike dictionaries, which provide all possible definitions of a word, glossaries present only context-relevant definitions. For example, when providing the definition of a person who is *rash*, only "careless and unwise" is indicated, whereas the definition related to the "reddening of a person's skin" is omitted (Cambridge Dictionary, n.d.). A meta-analysis by Pennock-Roman and Rivera (2011) on the effectiveness of several testing accommodations found that glossary provision was the most effective accommodation for increasing English language learners' comprehension of written material in English.

Although most research on glossary provision (e.g., Abedi et al., 2004) supports its use, a few studies using pop-up glossaries (i.e., glossaries on computerized tests that individuals access by clicking on designated spots) have questioned their utility. For example, Cohen and colleagues (2017) found that pop-up glossaries not only failed to improve performance on language arts and mathematics tests for Grade 3 students with non-native English speaking proficiency, but it also slightly negatively impacted the mathematics performance of Grade 7 students with non-native English speaking proficiency. Although there is also some evidence to indicate that a pop-up glossary can aid the mathematics and language arts performance of students with non-native English language proficiency, it also positively impacted the scores of native English-speaking students (Abedi et al., 2001). This finding is problematic because it suggests that glossary provision may have changed the construct validity of the test, meaning that the test may no longer have assessed what it was intended to measure.

There are two main limitations to using pop-up glossaries. First, the effort required to click on the designated spot to reveal definitions can reduce cognitive and time resources during testing, which can impact test performance. Second, the provision of pop-up glossaries does not guarantee that they are used by individuals. To address these limitations, we opted to use a within-text glossary to display definitions of potentially difficult words and phrases in brackets within test items (i.e., personality statements), thus making the process of reading key definitions relatively automatic.

Given the mixed findings on the effectiveness of glossary provision for standardized achievement testing and the lack of research on language proficiency, we sought to examine the influence of language proficiency and glossary provision on the validity and reliability of personality test scores. Using an experimental design in which participants with limited and high English proficiency were randomly assigned to a glossary condition, we gathered important evidence for determining the appropriateness of this potential testing accommodation. As Lovett and Lewandowski (2015) explained, a testing accommodation requires evidence for its differential benefit (i.e., test scores improve only for those who need them) and for unchanged test score inferences (e.g., unchanged validity, reliability). In this study we focus on establishing evidence for the latter, because unlike standardized achievement or cognitive ability tests in which higher scores reflect successful performance, such interpretations are untenable for personality test scores.

METHOD

Sample

Amazon Mechanical Turk (MTurk) and Prime Panels concierge online crowdsourcing services were used to recruit participants for this study. Of the 860 individuals from the United States and Canada invited to complete the screening questionnaire, 206 qualified for this study. Most individuals did not meet at least one of the study requirements which included (a) being at least 18 years old, (b) living in Canada or the United States, or (c) if scoring below eight on the English proficiency test (described later), being a non-native English speaker with one of 25 pre-identified languages as their native language (see Procedures section for more information). In addition, some individuals were excluded for failing attention checks (e.g., "please select agree" embedded within test items), providing the same response for all test items, or for completing the study in less time than was possible (i.e., less than 3 minutes). The majority of the participants were from the United States (96.6%) and were female (64.2%). This study included individuals of varying ages (23.3% [18-29 years], 39.3% [30-40 years], 30.6% [41-55 years], and 6.8% [56+ years]) and educational background (5.5% [less than a high school degree], 18.0% [high school degree], 13.0% [trade/vocational/technical degree], 11.5% [associate degree], 34.0% [bachelor's degree], 14.4% [master's degree], and 3.5% [advanced degree]). Most participants identified as Caucasian (41.2%) or Hispanic/Latino (40.7%), whereas the remaining participants identified as East Asian (11.3%) or other (6.9%). Participants were paid $1.80 US through MTurk, and an undisclosed amount from the Prime Panels concierge service fees as compensation.

Measures

English proficiency test. A 10-item multiple-choice English proficiency test (α = .77) was adapted from Powertutorials (2019) by replacing Question 1 with another item to improve clarity. A sample of items as well as estimates of each item's difficulty (indicated by the percentage of individuals answering each item correctly, which is generated by Powertutorials following completion of the test) is provided in Appendix A. The questions required individuals to choose the most appropriate word to complete sentences.

OCEAN. This short measure of personality of the Big Five (O'Keefe et al., 2012) was used to assess conscientiousness and neuroticism dimensions (4 items each; α = .85 and .75, respectively). Participants used a 7-point Likert scale to indicate the extent to which each item was characteristic of them.

International Personality Item Pool (IPIP) scales. The IPIP is a validated inventory with more than 3,000 personality items (Goldberg et al., 2006). From this pool, five conscientiousness (α = .81) and five neuroticism (α = .69) items were selected for use. Participants rated the extent to which each item represented them on a 5-point Likert scale.

HEXACO. The 10 conscientiousness (α = .79) and 10 neuroticism/emotionality (α = .78) items from the 60-item HEXACO (Ashton & Lee, 2009) were used in this study. Participants indicated the degree to which each item described them using a 5-point Likert scale.

Perceived usefulness. Two items, one for participants in the glossary group ("I found the definitions provided in brackets useful as they helped me understand the statements") and another for those in the no-glossary group ("I had difficulty understanding some of the words in the questionnaire. I wish definitions were provided for these words"), were developed for this study. These items were rated using a 5-point Likert scale from 1 = *strongly disagree* to 5 = *strongly agree*.

Procedure

Participants who answered eight or more questions correctly on the English Proficiency Test were classified as having high English proficiency, whereas all others were classified as having limited English proficiency. The limited English proficiency group also had to self-identify as a non-native English speaker and to speak one of 25 languages other than English. These languages were those in which the HEXACO was available for completion. As there was no pre-established cut-off for classifying individuals as having high or limited English language proficiency on the basis of this particular test, the cut-off decision was based on (a) the fact that a large proportion of the sample pool scored high on this measure (i.e., a ceiling effect) and (b) the need to have an equal number of participants in the high and limited language proficiency groups. The resulting average score on the English proficiency test for the limited (M = 4.37, SD = 1.73, n = 100) and high (M = 8.64, SD = 0.61, n = 106) language proficiency groups were significantly different, Welch's t (121.68) = -23.42, p < .01. In addition, using time data (recorded within the Qualtrics survey platform and calculated by taking the difference between the first and last click of each section containing the OCEAN and IPIP items, averaged across both measures), the limited English language proficiency group was found to take a significantly longer time in seconds (M = 92.64, SD = 86.25) than the high language proficiency group (M = 64.14, SD = 70.95, t = 2.55, p < .05) to complete these measures. These findings provide some support for the appropriateness of our classification of individuals on the basis of their scores on the English proficiency test.

Participants who were classified as having limited English proficiency were then presented with a list of the 25 languages in which the HEXACO was available and were asked to indicate if they were able to read and understand any of the languages without difficulty. Only those who selected one of the 25 languages were permitted to continue. All participants were then randomly assigned to the glossary or no-glossary condition. All participants completed subscales for two dimensions (conscientiousness and neuroticism) from the three personality measures described in the Measures section. These dimensions were chosen because of their generally higher validity in predicting job performance (Barrick & Mount, 2001). Participants in the high English proficiency group completed all measures in English, whereas those in the limited English proficiency group completed the HEXACO in a language of their choice and the other two measures in English. For those in the glossary condition, items on the IPIP and OCEAN measures were modified to include, in brackets, definitions of words and phrases that were potentially difficult to understand (as determined by the authors' collective input). Following completion of the personality measures, participants were asked about their perceived usefulness of (glossary condition) or perceived need for (no-glossary condition) definitions. A translated attention check item was used in the non-English HEXACO measures to ensure effortful responding and to confirm that participants were fluent in their claimed native language.

RESULTS

Table 1 reports descriptive statistics and correlations of the key variables of interest. Our central research questions pertained to the influence of language proficiency and glossary provision on the validity and reliability of personality test scores. We report on these findings below.

Construct Validity

As the HEXACO was completed in participants' native language, it was assumed to represent a "truer" measure of personality because it was uncontaminated by construct irrelevant variance due to test-taker language. Consequently, its relationship with the other personality measures (i.e., OCEAN and IPIP, which were completed in English) provided insight into the latter two measures' construct validity. We used moderated regression analyses to explore the influence of language proficiency and glossary provision on the construct validity of personality by entering the respective personality dimensions (i.e., conscientiousness, neuroticism) of the OCEAN/IPIP as independent variables, the HEXACO (respective dimensions) as the dependent variable, and language proficiency scores and glossary condition as moderators. These results are presented in Table 2. With respect to the influence of language proficiency, significant two-way interactions between language proficiency and both the personality measures/dimensions (i.e., OCEAN and IPIP/ conscientiousness and neuroticism) suggested that language proficiency affected the construct validity of scores on these personality measures.

With respect to glossary provision, we were interested in knowing whether this type of accommodation affected the validity of the personality measures. As reported in Table 2, the nonsignificant, two-way interactions between the OCEAN/IPIP and glossary condition for both personality dimensions suggested that the provision of a glossary during completion of the OCEAN/IPIP measures did not affect their convergent validity. A visual depiction (see Figures 1 and 2) of construct validities within each exper-

TABLE 1.
Descriptive Statistics and Correlations for Key Variables

	M	*SD*	1	2	3	4	5	6	7	8
1. Prof. score	6.57	2.49								
2. Prof. group[a]	1.51	0.50	.86**							
3. Glossary group[b]	1.50	0.50	.02	.01						
4. OCEAN_C	4.91	1.41	.15*	.12	.12					
5. OCEAN_N	3.90	1.45	-.07	-.09	-.01	-.02				
6. IPIP_C	3.57	0.64	.37**	.34**	.08	.60**	-.07			
7. IPIP_N	3.31	0.55	.27**	.22**	-.01	.06	.52**	.27**		
8. HEXACO_C	3.63	0.62	.12	.13	.12	.53**	-.12	.69**	.04	
9. HEXACO_N	3.34	0.64	-.07	-.03	.07	.03	.46**	.08	.58**	.05

Note. *N* = 206; [a] 1= limited English language proficiency and 2 = high English language proficiency; [b] 1= glossary provided and 2 = glossary not provided; Prof. = English language proficiency; OCEAN_C = Conscientiousness (OCEAN measure); OCEAN_N= Neuroticism (OCEAN measure); IPIP_C = Conscientiousness (International Personality Item Pool measure); IPIP_N = Neuroticism (International Personality Item Pool measure); HEXACO_C= Conscientiousness (HEXACO measure); HEXACO_N = Neuroticism (HEXACO Measure).
* *p* < .05. ***p* < .01

TABLE 2.
Regressions With OCEAN.20, IPIP, Language Proficiency, and Glossary Condition Predicting HEXACO Scores

Dimension	Predictor	*t*	β	*F* change	ΔR^2
OCEAN.20					
Conscientiousness	Step 1			27.14***	.29
	OCEAN.20	8.71***	.52		
	Proficiency	0.74	.04		
	Glossary	0.86	.05		
	Step 2			9.41***	.09
	OCEAN.20 x Proficiency	4.67***	.27		
	OCEAN.20 x Glossary	.19	.03		
	Glossary x Proficiency	-3.00**	-.55		
	Step 3			0.01	.00
	OCEAN.20 x Proficiency x Glossary	-.08	-.01		
Neuroticism	Step 1			18.96***	.22
	OCEAN.20	7.39***	.46		
	Proficiency	-.60	-.04		
	Glossary	1.14	.07		
	Step 2			4.71**	.05
	OCEAN.20 x Proficiency	3.41**	.21		
	OCEAN.20 x Glossary	.36	.07		
	Glossary x Proficiency	1.08	.21		
	Step 3			0.51	.00
	OCEAN.20 x Proficiency x Glossary	.71	.15		
IPIP					
Conscientiousness	Step 1			65.81***	.49
	IPIP	13.65***	.74		
	Proficiency	-2.81**	-.15		
	Glossary	-1.14	.06		
	Step 2			6.81***	.05
	IPIP x Proficiency	3.57***	.18		
	IPIP x Glossary	-.57	-.10		
	Glossary x Proficiency	-2.51*	-.43		
	Step 3			2.38	.01
	IPIP x Proficiency x Glossary	-1.54	-.25		

continued

TABLE 2. (CONTINUED)

Regressions With OCEAN.20, IPIP, Language Proficiency, and Glossary Condition Predicting HEXACO Scores

Dimension	*Predictor*	*t*	β	*F* change	ΔR^2
Neuroticism	Step 1			44.14***	.40
	IPIP	11.38***	.65		
	Proficiency	-4.22***	-.24		
	Glossary	1.39	.08		
	Step 2			6.88***	.06
	IPIP x Proficiency	4.27***	.23		
	IPIP x Glossary	-.32	-.06		
	Glossary x Proficiency	1.47	.26		
	Step 3			0.69	.00
	IPIP x Proficiency x Glossary	.83	.15		

Note. Proficiency coded *limited* = 0 and *high* = 1; Glossary coded *glossary* = 1 and *no-glossary* = 0; *** *p* < .001; ** *p* < .01

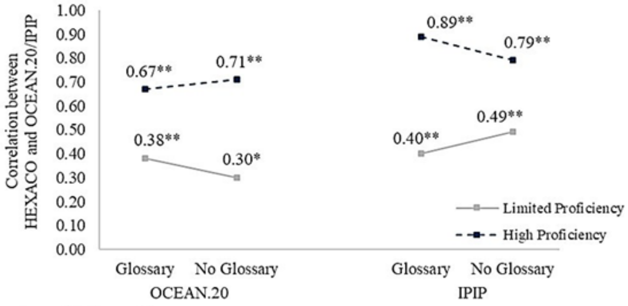
imental condition further shows that the respective associations (i.e., between the HEXACO and OCEAN/IPIP conscientiousness/ neuroticism dimensions) no-glossary were uniformly similar and did not significantly differ from each other (*p* > .05). This finding provides some evidence for the appropriateness of within-text glossaries as an accommodation, in that the accommodation did not affect the measures' construct validity, allowing for similar test score inferences under standardized (i.e., no-glossary) and accommodated (i.e., glossary) test administration conditions. Table 2 further shows nonsignificant, three-way interaction effects, suggesting that the validities of the OCEAN/IPIP across the two glossary conditions do not vary with language proficiency. Tests of significance comparing these associations within each language proficiency group (as shown in Figures 1 and 2) were also nonsignificant (*p* > .05). Moderated regression analyses within each language proficiency group also yielded nonsignificant interaction effects (glossary x personality measure) for each personality measure (*p* > .05).

Reliability

With respect to reliability, on one hand, we were looking for evidence (i.e., no differences in reliability across conditions) to support Lovett and Lewandowski's (2015) unchanged construct requirement for the appropriateness of a test accommodation. However, given the specific nature of the test accommodation (i.e., glossary provision), it was also plausible to expect the provision of a glossary to compensate for language-related difficulties experienced by individuals with limited English proficiency, thereby resulting in scores that contain less measurement error (or higher reliability). In other words, personality test scores of limited English proficiency participants who received a glossary should be more reliable than those who did not receive a glossary.

FIGURE 1.

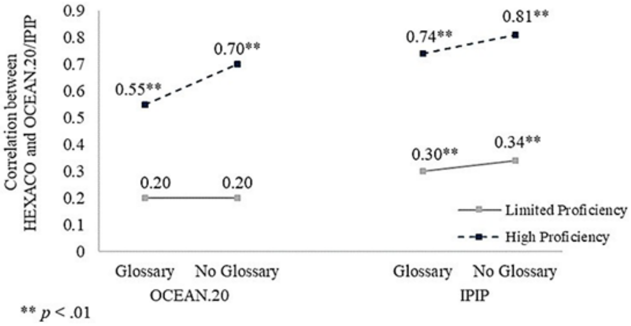
Conscientiousness Correlations Between the OCEAN.20/IPIP and HEXACO



* *p* < .05, ** *p* < .01

FIGURE 2.

Neuroticism Correlations Between the OCEAN.20/IPIP and the HEXACO



** *p* < .01

We first examined estimates of reliability (i.e., Cronbach's alpha) of the OCEAN/IPIP personality dimensions across the two glossary conditions. As presented in Table 3, the coefficients appear to be similar across conditions. Using Feldt's (1969) *F*-test for the comparison of two independent alpha coefficients (*cocron* package in R; Diedenhofen, 2016), we further confirmed that these reliability estimates did not differ ($p > .05$) across glossary conditions for both personality measures and dimensions, providing further support for unchanged test score inferences when a within-text glossary is provided during completion of these measures. To examine whether this finding depended on language proficiency, we examined Cronbach's alpha values across glossary conditions within each language proficiency group (see Table 3). Focusing on the limited language proficiency group, we statistically compared coefficients across glossary conditions to find that none of the coefficients differed significantly across the glossary conditions (p > .05), although the finding for the OCEAN conscientiousness dimension approached significance ($F(49, 49) = 1.73$, $p = .057$).

Despite the nonsignificant differences between coefficient alphas, a visual inspection of these coefficients (see Table 3) showed that all coefficients in the high language proficiency group were above the conventional acceptable value of .70 (Cortina, 1993), whereas four coefficients in the low language proficiency group were below this value. Consequently, we were interested in exploring any consequential effects of such deviations from conventional reliability values. We computed the standard error of measurement (SEM) to estimate the potential gain/loss in measurement precision (Harvill, 1991). Table 4 reports these values and shows that for the limited proficiency group, the provision of a glossary resulted in a 33% net loss in measurement precision (for the OCEAN conscientiousness dimension) but a nearly 11% gain in measurement precision for the neuroticism dimension of this same measure. Therefore, although the reliability coefficients for the OCEAN dimensions were found not to differ significantly across glossary conditions, the resulting amount of change in measurement precision across these conditions was not negligible. The direction of change, however, was inconsistent across the two personality dimensions assessed by the OCEAN, including being counter to what we expected for conscientiousness.

Usefulness of Glossary Provision

We compared the perceived usefulness of the glossary across the limited and high English proficiency groups (see Table 5). Recall that participants in the glossary condition were asked whether they found the glossary useful, whereas those in the no-glossary condition were asked if they thought a glossary would have been useful. There was a small statistically significant difference in usefulness ratings between limited and high English proficiency participants who were provided a glossary, $F(1,100) = 4.87$, $p = .03$, $\eta^2 = .05$, but it was counter to what was expected, as high proficiency participants found the glossary more useful. For those in the no-glossary condition, there was a moderate statistically significant difference in ratings between the two language proficiency groups, with those in the high English proficiency group tending to disagree more strongly ($M = 1.28$, $SD = 0.72$) than those in the limited proficiency group ($M = 2.63$, $SD = 1.30$) that the words were difficult and that definitions were required ($F(1,100) = 42.89$, $p < .001$, $\eta^2 = .30$).

TABLE 3.

Cronbach's Alpha for Conscientiousness and Neuroticism for Each Glossary Group

		Conscientiousness		Neuroticism	
English proficiency	Measure	Glossary	No glossary	Glossary	No glossary
Limited	OCEAN.20	.74	.85	.76	.58
	IPIP	.56	.72	.62	.51
High	OCEAN.20	.92	.90	.75	.82
	IPIP	.86	.85	.73	.78

Note. For limited English proficiency participants, $n = 50$ for each of the glossary groups. For high English proficiency participants, $n = 52$ and 54 for the glossary and no-glossary groups, respectively.

TABLE 4.
Standard Error of Measurement (SEM) and Net Gain/Loss in Measurement Precision

		Conscientiousness			Neuroticism		
English proficiency	Measure	Glossary	No glossary	% Net diff	Glossary	No glossary	% Net diff
Limited	OCEAN.20	.77	.58	*-33.03*	.72	.80	*10.56*
	IPIP	.32	.32	*0.00*	.32	.32	*0.00*
High	OCEAN.20	.39	.39	*0.00*	.71	.70	*-2.04*
	IPIP	.25	.25	*-1.14*	.29	.30	*3.29*

Note. Net diff = Net difference. Net diff values are such that positive values indicate a net gain in measurement accuracy and negative values indicate a net loss. SEM values were not rounded in net difference calculations.

TABLE 5.
Mean Usefulness Ratings With Standard Deviations

English proficiency	Glossary *M* (*SD*)	No-glossary *M* (*SD*)
High	4.17 (0.98)	1.28 (0.72)
Limited	3.72 (1.08)	2.63 (1.30)

Note. Limited English proficiency participants: *n* = 50 for each of the conditions. For high English proficiency participants: *n* = 52 and 54 for the glossary and no-glossary groups, respectively.

DISCUSSION

Practical and Theoretical Contributions

To our knowledge, this is one of the first studies to examine the influence of language proficiency and glossary provision on personality measurement. The appropriateness of glossaries as a potential accommodation during personality testing must be determined. Although accommodations during testing are supported by standard selection guidelines (e.g., *Principles for the Validation and Use of Personnel Selection Procedures*, Society for Industrial and Organizational Psychology [SIOP], 2018), there is a requirement to document evidence supporting its use. In particular, evidence must show that the construct(s) measured by the test do(es) not change and that comparable inferences can be made from test scores (Lovett & Lewandowski, 2015; Phillips, 1994). Using an experimental design in which participants in low and high language proficiency groups were assigned to a glossary condition, we set out to do exactly that.

Our findings on the convergent validity of the OCEAN.20 and IPIP measures of conscientiousness and neuroticism with respective dimensions of the HEXACO, completed in participants' native language, provided evidence that the construct validity of these measures did not change when a glossary was provided. In addition, these findings did not depend on English language proficiency. We found similar support when we examined reliability; Cronbach's alpha coefficients were found not to differ significantly across glossary conditions, even when examined within each language proficiency group. However, the net change in measurement precision within the limited language proficiency warrants further examination, especially because they were not consistent across measures. We also obtained insight into the perceived usefulness of a within-text glossary to participants. Both limited and high language proficiency participants who were provided with a glossary found the provision to be at least somewhat useful, but when it was not provided neither group expressed a need to have it. These findings suggest that the test items may have been sufficiently easy enough to understand without the provision of a glossary for both groups of individuals. Overall, the findings from this study provide important insight into the influence of language proficiency and the use of glossaries on personality measurement. It also raises an awareness about the need for and nature of evidence required to support the appropriateness of any accommodation during testing.

Limitations and Future Research Directions

This research has three potential limitations, some of which highlight avenues for further research in this area. First, we acknowledge that the English proficiency test and cut-off scores used to assign participants to the high or limited proficiency group may not have adequately classified individuals into the two groups. This was partly a result of our screening criteria (i.e., stipulating that a person had to reside in North America), which resulted in a high number of participants scoring high on this measure. However, as reported earlier, the average scores on the English proficiency test within each language proficiency group were significantly different from each other, and the groups also differed in the amount of time they took to complete the personality measure, providing some assurance that the groups were appropriately classified. Nevertheless future research should use a more well-established test with validated cut-off scores to classify participants according to their level of English proficiency.

Second, one of the assumptions of this study was that all individuals who were provided with a glossary read the definitions because they were placed beside key words and phrases. However, to help confirm this assumption, we examined the completion times for each personality measure (i.e., OCEAN and IPIP) and found that, although nonsignificant, the average time in seconds was higher in the glossary condition (M_{OCEAN} = 67.01, SD_{OCEAN} = 72.67; M_{IPIP} = 99.81, SD_{IPIP} = 95.57) than in the no-glossary condition (M_{OCEAN} = 60.99, SD_{OCEAN} = 78.77; M_{IPIP} = 84.46, SD_{IPIP} = 103.34). These differences were, however, more pronounced and significant in the limited language proficiency group; glossary condition (M_{OCEAN} = 89.87, SD_{OCEAN}= 96.98; M_{IPIP} = 125.82, SD_{IPIP} = 124.83) and no-glossary condition (M_{OCEAN} = 60.46, SD_{OCEAN}= 36.66; M_{IPIP} = 94.41, SD_{IPIP}= 86.52). The differences were significant for both measures (OCEAN: Cohen's d = .40, t = 2.84, p < .05; IPIP: Cohen's d = .29, t = 2.07, p < .05). Given these findings, it is plausible that participants paid some attention to the within-text definitions.

Another possibility is that the within-text placement of definitions may have increased cognitive load. Future studies may explore different survey designs (e.g., pop-up boxes) to better assess whether provided definitions were read. A pop-up glossary involves providing participants with the definitions of potentially difficult words or phrases through pop-up windows that can be accessed by hovering over the potentially difficult words or phrases (see Cohen et al., 2017). This type of glossary may be associated with reduced cognitive load as participants can choose if they want to view definitions. The reduced cognitive load may lead to more valid personality responses as participants can focus their cognitive effort on interpreting the most important information. Thus, future research should examine the effect of a pop-up glossary on the validity and reliability of personality responses.

Third, future research can benefit from larger sample sizes and a more detailed language proficiency classification system to comprehensively examine the potential benefits of glossary provision. For example, having a sufficient sample size to incorporate a "moderate" language proficiency group would have helped to examine if glossary provision, although not beneficial for individuals with limited or high proficiency, may have been useful for those with moderate levels of English proficiency.

Concluding Remarks

Despite its limitations, this study is one of the first to examine the influence of language proficiency and the provision of glossaries during personality testing. It sheds light on issues pertaining to the use of a glossary accommodation and the need to obtain evidence to support the appropriateness of a glossary accommodation, including the need to obtain evidence to evaluate its impact on the measurement properties of a personality test.

REFERENCES

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). NAEP math performance and test accommodations: Interactions with student language background. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. Review of Educational Research, 74(1), 1–28.

Abedi, J., Leon, S., & Mirocha, J. (2003). Impact of student language background on content-based performance: Analyses of extant data (CSE Technical Report 603). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Abedi, J., Lord, C., & Plummer, J. R. (1997). Final report of language background as a variable in NAEP mathematics performance. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. Journal of Personality Assessment, 91(4), 340–345. https://doi.org/10.1080/00223890902935878

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? International Journal of Selection and Assessment, 9(1–2), 9–30. https://doi.org/10.1111/1468-2389.00160

Butler, F. A., & Castellon-Wellington, M. (2005). Students' concurrent performance on tests of English language proficiency and academic achievement. In J. Abedi, A. Bailey, F. Butler, M. Castellon-Wellington, S. Leon, & J. Mirocha (Eds.), The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives, 47–83. National Center for Research on Evaluation, Standards, and Student Testing.

Cambridge Dictionary. (n.d.). Rash. In Cambridge Dictionary. Retrieved from https://dictionary.cambridge.org.

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. Linguistic and Cultural Influences on Learning Mathematics, 17–46.

Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. Applied Measurement in Education, 30(4), 259–272.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78(1), 98.

Diedenhofen, B. (2016). Package cocron: Statistical comparisons of two or more alpha coefficients. https://mirror.rcg.sfu.ca/mirror/CRAN/web/packages/cocron/cocron.pdf

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84-96.

Harvill, L. M. (1991). Standard error of measurement: An NCME instructional module on. Educational Measurement: Issues and Practice, 10(2), 33–41.

John O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. European Journal of Personality, 2, 171-203.

Lovett, B. J., & Lewandowski, L. J. (2015). Testing accommodations for students with disabilities: Research-based practice. American Psychological Association.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. Harvard Educational Review, 78(2), 333–368.

O'Keefe, D. F., Kelloway, E. K., & Francis, R. (2012). Introducing the OCEAN. 20: A 20-item five-factor personality measure based on the trait self-descriptive inventory. Military Psychology, 24(5), 433–460.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. Educational Measurement: Issues and Practice, 30(3), 10–28.

Phillips, S. E. (1994). High–stakes testing accommodations: Validity versus disabled rights. Applied Measurement in Education, 7, 93–120. https://doi.org/10.1207/s15324818ame0702_1

Powertutorials. (2019, June 11). Quick English level test [web quiz]. Retrieved from https://www.proprofs.com/quiz-schoo1/story.php?title=quick-english-1evel-test

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures (5th ed.). Cambridge University Press.

Umbel, V. M., Pearson, B. Z., Fernández, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. Child Development, 63(4), 1012–1020.

United States Census Bureau. (2000). Table 3: Language use, English ability, and linguistic isolation for the population 18 years and over by State: 2000. https://www2.census.gov/programs-surveys/decennial/2000/phc/phc-t-10/tab03.pdf

Verhallen, M., & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. Applied Linguistics, 14(4), 344–363.

RECEIVED 08/12/2021 ACCEPTED 02/15/2022

Appendix A

Sample Items From the English Proficiency Test (Power Tutorials, 2019)

1. Can you hear what he is …?
   A. saying
   B. speaking
   C. telling
   D. talking

2. She hasn't come home …
   A. still
   B. till
   C. yet
   D. already

Percentage (%) of Test Takers who Answered Each Question (Q) Correctly

	English proficiency test questions										
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Average
Powertutorials (2019)	86	91	96	89	62	76	55	35	N/A	69	73.2
All participants in this study	75	85	90	17	66	68	66	61	55	73	65.6
Low proficiency participants	49	69	79	22	36	37	45	27	25	48	43.7
High proficiency participants	100	100	100	13	93	98	85	93	84	97	86.3