Open Journal of Modern Linguistics
2012. Vol.2, No.1, 34-41
Published Online March 2012 in SciRes (http://www.SciRP.org/journal/ojml)
http://dx.doi.org/10.4236/ojml.2012.21005

A Preliminary Version of an Internet-Based Picture Naming Test

Anatoliy V. Kharkhurin
Department of International Studies, American University of Sharjah, Sharjah, UAE
Email: akharkhurin@aus.edu

Received December 16th, 2011; revised February 15th, 2012; accepted February 22nd, 2012

The study presents a web-based productive vocabulary assessment tool, the internet Picture Naming Test (iPNT). The iPNT is administered online and takes eight minutes to complete. The iPNT assesses vocabulary knowledge by rating participants' responses to 120 colored drawings of simple objects. Participants type the names of the objects and the names are saved as a computer file that can be uploaded into statistical software for further processing. The test is rated by comparing participants' responses against a list of correct labels. High test-retest reliability suggests that iPNT can be considered a reliable measure. The study evaluates convergent validity of the iPNT by comparing its scores with paper-based and oral versions of the same test and concurrent validity by comparing its scores with that of receptive Peabody Picture Vocabulary Test, a language aptitude Cloze test, standard admission Test of English as a Foreign Language, and DIALANG diagnostic tool. Highly significant correlations between the scores on these tests and iPNT scores suggest that the latter is a suitable assessment tool for language proficiency. However, the moderate correlation values ranging from .52 to .68 indicate that the use of this test should be limited to psychometric research assessing an individual's productive vocabulary knowledge.

Keywords: Online; Language Proficiency; Picture Naming; Test

Introduction

The realities of the contemporary world with its vast migrations, massive international cultural, political and economic interactions encouraged the representatives of different national, ethnic, religious and social groups to select a common language that facilitates mutual communication. The contemporary *lingua franca*, English, became the vehicular language in various areas of human endeavor, including academic communities. With the increasing number of international academic communities requiring English as second language, there is a growing demand in testing of language skills. Most educational institutions have an English exam as an admission criterion for non-native speakers (e.g., Test of English as a Foreign Language, TOEFL; Cambridge English for Speakers of Other Languages certificates). These tests present a comprehensive assessment of a variety of linguistic skills (e.g., internet-based TOEFL assesses listening, writing, reading, and speaking skills). The European Community sponsored an on-line diagnostic language assessment system DIALANG[1], which is based on the Common European Framework of Reference (Council of Europe, 2001). It provides learners with information about their language proficiency in 14 European languages and informs them of their Common European Framework level (Chapelle, 2006). This tool assesses the learner's skills in listening, writing, reading, structure, and vocabulary.

In addition to educational purposes, testing of linguistic skills becomes an important issue for empirical research in first and second language acquisition and bilingualism. A large portion of studies in these fields either lack control over participants' language proficiency (see Kharkhurin, 2005, for an overview) or employ admission tests developed for educational purposes. However, one apparent problem with using the admission tests is the duration of the testing, which cannot be suitable for experimental conditions. For example, the internet-based TOEFL may take around four hours (ETS, 2008b) and DIALANG may last for up to two hours. Although, this time frame is feasible in a class-room setting, it may become a challenge for time constrained experimental settings with adults. Therefore, contemporary language related empirical research focuses on increasing the number of tests that can be completed within a relatively short time interval. These tests use different techniques and claim to assess different linguistic skills.

Several techniques have gained a reputation of being a reliable language assessment tool and have thus been widely used in empirical research. The Cloze procedure (Taylor, 1953) asks participants to complete written text with various gaps. Extensive empirical investigation presents such tests as assessments of overall native language proficiency (e.g., Dupuis, 1980; Peterson, Peters, & Paradis, 1972). For example, Dupuis found a Cloze test to be a good predictor of reading comprehension in monolingual 10th graders. This test was also shown as a reliable indicator of foreign language proficiency in second language learners (e.g., Baldauf & Propst, 1979; Oller, 1972, 1973; Oller & Conrad, 1971). It was found to strongly correlate with the TOEFL (Darnell, 1968). Another study demonstrated the ability of a Cloze test to discriminate between English native speakers who learned German, Japanese, Russian, or Spanish for the first, second, or third semester, i.e. the scores on the Cloze test correlated with the length of foreign language study (Briere, Clausing, Senko, & Purcell, 1978). Jochems and Montens (1987) tested second language Dutch learners on a Cloze test and four tests of language proficiency: listening, speaking, reading, and writing, conducted by the national workgroup *Centrale Toets Nederlands*. They found that the scores on their Cloze test highly correlated with all four tests of language proficiency and appeared to form a solid basis for prediction of the

[1]http://www.lancs.ac.uk/researchenterprise/dialang/about.htm

total scores for all these tests taken together.

A second widely used test, the Peabody Picture Vocabulary Test (PPVT) of receptive vocabulary, asks participants to indicate which of the four shown pictures corresponds to a name spoken by the experimenter. Clinicians and researchers rely on the test to accurately assess children's and adults' single word lexical knowledge. According to the user's manual of the fourth edition of the test (Dunn & Dunn, 2007), PPVT scores correlated with the scores of the second edition of the Expressive Vocabulary Test (Williams, 2007; mean r = .82 across age groups), the Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999; adjusted r ranging from .41 to .79 for various test activities and age groups), the fourth edition of a more comprehensive Clinical Evaluation of Language Fundamentals (Semel, Wiig, & Secord, 2003; adjusted r ranging from .67 to .75 for various test activities and age groups), and the Group Reading Assessment and Diagnostic Evaluation (Williams, 2001; mean r = .63 across various test levels).

A third commonly used test of productive vocabulary, the Boston Naming Test (BNT, Kaplan, Goodglass, & Weintraub, 1983) is generally used by clinicians to assess word retrieval performance of brain-damaged patients. The test consists of 60 outline drawings of objects and animals presented in the order of word frequency and grade of difficulty. Participants are asked to name the pictures arranged in a booklet. This test was recognized as a reliable tool to identify naming deficits and impaired word-retrieval capacities in a variety of cerebral pathologies in an adult and a childhood population (see Mariën, Mampaey, Vervaet, Saerens, & De Deyn, 1998, for a summary). It has been translated into several languages and administered to healthy populations from a variety of age and gender groups with different educational backgrounds in various geographic regions (see Patricacou, Psallida, Pring, & Dipper, 2007, for a summary).

In contrast to TOEFL and DIALANG, the Cloze test, PPVT, and BNT require considerably less administration time. However, they still need to be administered and rated by an experimenter. The purpose of the present study is to present a new measure that provides a fully automatic, rater independent, and reliable measure of language proficiency that can be administered in any location in a relatively short time interval. The test employs a technique similar to the BNT; that is, participants are presented with drawings of objects and asked to name these objects. There are four major differences of the proposed picture naming test (PNT) and the original BNT. First, it is administered on the web and therefore can be accessed worldwide. Second, the responses are to be provided in a written form (not orally as in the BNT), which eliminates the need for the experimenter to be present. Third, it is timed and therefore it ensures equal testing time for all participants. This condition appears to be crucial when the test is administered in an uncontrolled manner. Fourth, the written responses are recorded in a computer file that can be automatically uploaded to statistical software for further analysis.

An obvious limitation of this test, as well as many other language proficiency tests used in psycholinguistic research, is its inability to assess all four major language skills: speaking, writing, listening, and reading (cf. Padilla & Ruiz, 1973). However, the limited testing scope is compensated for by a short testing time. A high predictive power of these tests may also reconcile the researchers with this limitation. Indeed, Kharkhurin (2005) found in a pilot study that paper-based written PNTs in English and Russian highly correlated with the Cloze procedure in these respective languages (r = .77, p < .01 for English; and r = .83, p < .01 for Russian). Another study found this test to strongly correlate with participants' self-rating of language skills in English and Russian and their self-assessment of the degree of Russian-English bilingualism (Kharkhurin, 2008). To ensure the concurrent validity of the PNT in the present study, participants' performance on this test was compared with a battery of other measures including an assessment tool (TOEFL), a diagnostic tool (DIALANG), a common test of language proficiency (a Cloze procedure), and a widely used test in psycholinguistic research (the PPVT). To ensure convergent validity of this test, in addition to the internet-based version (iPNT), the PNT was presented to the same individuals in paper-based written and oral forms.

The objective of the present study was to present the methodology of the web-based productive vocabulary test with three crucial characteristics: 1) iPNT is internet-based; 2) iPNT is timed; 3) iPNT automatically produces a statistical software compatible output file with an individual's responses. Experiment 1 of the study aimed to provide evidence for convergent validity of the iPNT by comparing performance on this test with that on other versions of the PNT, and concurrent validity of the iPNT by correlating performance on this test with that on TOEFL, DIALANG, a Cloze test, and the PPVT. Experiment 2 evaluated test-retest reliability of the iPNT by administering the test to the same group of participants twice.

Experiment 1

Method

Participants

The participants were 87 American University of Sharjah (United Arab Emirates) students (29 male and 58 female; aged between 17 and 34, M = 20.10, SD = 2.19) who were recruited from the General Psychology subject pool. Although, they varied in their countries of origins (representing Middle East, Asia, Africa, North America, and Europe) and therefore in the distribution of their native languages, all of them were fluent in English due to the fact that English is the language of instruction at the University.

Instruments and Procedure

Upon completion of an online biographical questionnaire and submitting a copy of the TOEFL, participants were given a battery of language proficiency tests that were distributed between two sessions. One session included Cloze procedure, DIALANG online testing, and internet- and paper-based versions of PNT. One of the PNT versions was presented at the beginning and the other at the end of the session to minimize the priming effect. The other session included the PPVT and an oral version of the PNT. The presentation order of the tests in both sessions was counterbalanced across participants.

Biographical Questionnaire

An online multilingual and multicultural experience questionnaire[2] was administered to determine participants' linguistic and cultural background. They received a questionnaire that among other issues, obtained data on each participant's place of origin, languages they speak, their assessment of linguistic

[2]http://surveys.aus.edu/index.php?sid=87644

skills in each of these languages, and age of acquisition of these languages.

Picture Naming Test

This test was initially designed by Kharkhurin (2005) as a test of productive vocabulary, which assesses language proficiency as the accuracy of participants' responses to pictures of simple objects, a technique similar to the BNT and the one used by Lemmon and Goggin (1989). The test stimuli are 120 pictures of simple objects (Appendix A) randomly selected from those scaled by Rossion and Pourtois (2004), an improved version of Snodgrass and Vanderwart (1980). The procedure requires participants to produce a name of the object presented in the picture, which they would normally use in everyday life.

Three versions of the PNT were used in the present study. In the *oral* version, each picture was presented separately on the computer screen using Microsoft PowerPoint's full screen mode. Participants were asked to label an object in the picture by saying its name out loud. There was no time restriction for this test, but participants were encouraged to respond as fast as possible. Responses were recorded using Microsoft Sound Recorder version 5.1 software and played back during rating. In the *paper-based* version, the pictures were arranged on four pages. Participants wrote down their responses in a booklet with numbered lines corresponding to the pictures. Each participant was given two minutes to label as many as possible of the 30 pictures on each page. In the *internet-based* version[3], the pictures are presented using LimeSurvey version 1.72 environment. They are arranged in four groups each of which appears on a separate webpage; each picture is accompanied by a 50 character space provided for an answer. The presentation order of the pictures within each group is randomized. Participants are given two minutes to label as many as possible of the 30 pictures on each page. The timer in the top left corner of the page indicates the elapsed time. After the time is elapsed, an "out of time" message appears on the screen and next page is loaded automatically.

The scoring procedure was the same for all three versions. Each response was scored either 1 or 0, so that the maximum number of points for picture naming was 120. A list of appropriate labels was generated for each picture. A list of primary labels was adopted from Snodgrass and Vanderwart (1980). The average name agreement coefficient[4] for these labels was .50 with 94.79% of participants giving the primary label, which according to Snodgrass and Vanderwart, suggests high name agreement for these labels. The word frequency for the primary list ranged from 1 to 431 (M = 35.01, SD = 70.40) per million according to Kučera and Francis (1967) and from .12 to 483.06 (M = 33.26, SD = 68.44) per million according to Brysbaert and New (2009). A list of secondary labels was formed based on the synonyms for the primary lables obtained by Snodgrass and Vanderwart. The word frequency for both primary and secondary lists ranged from .12 to 509.37 (M = 35.75, SD = 76.60) per milion acording to Brysbaert and New. If the participants' response matched the corresponding item on the list, they scored 1 point; otherwise, 0 points. Two sets of rating strategies were used: the *primary* rating gave a point only if the label from a primary list was used; the *secondary* rating gave a point if the label from either primary or secondary list was used; the *strict* rating gave a point if the produced label was spelled correctly; the *lenient* rating disregarded the spelling errors. Therefore, the paper- and internet-based PNTs received four scores: primary strict, primary lenient, secondary strict, and secondary lenient, and oral PNT received two scores: primary and secondary.

[3]http://surveys.aus.edu/index.php?sid=31316

[4]Snodgrass and Vanderwart (1980) defined a name agreement coefficient as a distribution of names given to a picture across participants. A picture that obtained the same name from every participant had a name agreement coefficient equal to .00 (perfect name agreement). A picture that obtained exactly two different names with equal frequency had a name agreement coefficient equal to 1.00. Increasing name agreement value indicated decreasing name agreement.

Cloze Procedure

The materials for the Cloze procedure were adopted from the practice tests for Cambridge English for Speakers of Other Languages certificates, an examination for people who use everyday written and spoken English at an upper-intermediate level for work or study purposes. Participants were asked to complete text with various linguistic gaps. In the rational selection procedure (Jongsma, 1980), the words of different lexical categories were deleted from the text fragments and substituted by blank spaces; participants had to insert the missing words. Two versions of the Cloze procedure were employed: a multiple-choice and an open-end acceptable response tasks. In the *multiple-choice Cloze task*, participants were asked to select one out of four words that best fits the blank space. In the *open-end acceptable response Cloze task*, they were asked to provide a word that best fits the blank space. Two texts with 15 blank spaces were supplied for each version of the test. Each participant received two texts, one of each version, preceded by a written instruction that explained the procedure and provided an example. The texts were presented to participants in a counterbalanced order to prevent any task version or fatigue effect. Participants had 10 minutes to complete both texts. They were given 1 point if their answer matched the one in the Cambridge English for Speakers of Other Languages certificates' answer key and 0 otherwise, which resulted in a maximum score for the Cloze procedure equating 30.

Peabody Picture Vocabulary Test IV, Form A

This is a standardized test of receptive vocabulary (Dunn & Dunn, 2007). The task is to indicate one out of four pictures shown on the test plate, which corresponds to the name given by the experimenter. The plates are arranged in order of increasing difficulty and grouped in 19 sets of 12 trials each. A raw score is calculated using basal and ceiling sets determined by the scoring procedure. The basal set is the easiest one in which a participant makes one error or less; the ceiling set is the one in which a participant makes eight errors or more. Testing continues until the ceiling set is determined. The raw score is calculated as a difference between a number of all possible correct responses (computed as a ceiling set number multiplied by 12) and a number of errors made by a participant during testing. The raw score is converted to a standard score by the recommended procedure, which takes age-related norms into account.

Test of English as a Foreign Language

The standard TOEFL certificate comes in three versions:

internet-based, computer-based, and paper-based. The reliability estimates for the test range from .74 to .85 for different skills and .94 for the total score (ETS, 2008a). Zhang (2008) compared the test scores of 12,385 examinees who have taken two internet-based TOEFLs within a period of one month. The correlations of their scores on the two test forms were .77 for listening and writing sections, .78 for reading, .84 for speaking, and .91 for the total test score.

The American University of Sharjah admission procedure requires all students to obtain a minimum of 71 for internet-based, 197 for computer-based, or 530 for paper-based TOEFL. This requirement ensured that all participants had their TOEFL certificate, and therefore they were asked to submit it before the beginning of the testing. The obtained scores from different certificate versions were converted into computer-based scores using TOEFL score comparison conversion tables (ETS, 2005). The computer-based scores for listening, reading, writing, and total were used in the further analyses. It is important to note however that different participants have taken the TOEFL examination at different times before the current testing (ranging from 1 to 7 years, M = 3.07, SD = 1.09), and therefore their scores cannon be considered an accurate measure of their language proficiency at the time of testing.

DIALANG

The English version of DIALANG was used in the present study. This testing system consists of a number of activities assessing the linguistic skills in five domains on three levels of difficulty (see Alderson & Huhta, 2005, for detailed description). In the beginning, participants are asked to do a Vocabulary Size Placement Test (VSPT), which is used to estimate the vocabulary size and to determine the level of subsequent testing. In the VSPT, participants have to decide whether the letter string presented is a word or a non-word (e.g., "to study" is a word, "to futt" is a non-word). The test uses 75 verbs (50 words and 25 non-words) presented in a random order, and the VSPT score range is 1 - 1000. After the placement test, participants are presented with five modules assessing linguistic skills in listening, writing, reading, structure, and vocabulary, which they can take in the order of their preference. The first three modules are preceded by a self-assessment questionnaire, in which participants are asked to make judgments about their abilities in the selected language skill by validating 18 statements per skill (e.g., for listening: "I can catch the main points in broadcasts on familiar topics and topics of personal interest when the language is relatively slow and clear."). The self-assessment is also used to determine the level of subsequent testing. After completion of both the VSPT and the self-assessment, the system combines the two results to decide which level of linguistic skill testing to administer. In the listening module, participants hear a short vocal presentation and receive questions based on this presentation. In the writing module, participants are asked to fill in the gaps in the text. In the reading module, they are asked to read a short text and answer questions based on this material. In the structure module, participants' knowledge of grammar is probed. Finally, the vocabulary module assesses participants' understanding of the words. The test items come in four different formats: multiple choice, drop-down menus, text entry and short-answer questions. All self-assessment and testing modules are scored according to six levels of the Common European Framework of Reference scale (Council of Europe, 2001) listed in the order of increased proficiency: A1, A2, B1, B2, C1, C2.

Results

The purpose of this experiment was to investigate the validity of the internet-based PNT. Therefore, participants' iPNT scores were compared with the paper-based and oral versions of the PNT as well as with the scores on the standardized measures of language proficiency: TOEFL, PPVT, Cloze, and DIALANG.

Picture Naming Test

The mean scores for all three versions of the PNT are presented in **Table 1** and the correlations between these scores are shown in **Table 2**. All four testing strategies applied to each of three testing modes obtained nearly perfectly correlated scores. The correlations between different PNT versions were also significantly high. These results suggest that various testing modes and the rating strategies assess the vocabulary knowledge similarly. This finding provides a justification for employing the internet-based PNT version that can be rated using the primary list of labels and strict spelling. The primary strict iPNT scores are used in the further analyses.

Note however that all three versions differ in the magnitude of obtained scores. Oral PNT rated with the primary list of labels obtained significantly higher scores than its internet- and paper-based counterparts rated with the lenient condition[5] (t = 4.73, $p < .001$ and t = 2.81, $p < .01$, respectively); the latter two scores were not significantly different. When the secondary list of labels was allowed, the oral PNT obtained the highest scores followed by paper-based (t = –2.68, $p < .01$) and internet-based PNTs (t = –7.39 and t = –3.77, respectively, both ps < .001). Participants were also found to obtain significantly higher scores on the paper-based PNT rated with the primary list of labels using the strict condition than on the iPNT rated with the same strategy (ΔM = 6.00, t = 4.66, $p < .001$).

TOEFL

Table 3 presents correlations for the computer-based TOEFL scores for listening, reading, writing, and total. The multiple regression analysis revealed that listening and reading scores

Table 1.
Mean scores and standard deviations (SD) for internet-based, paper-based, and oral PNTs obtained by applying four rating strategies: primary strict, secondary strict, primary lenient and secondary lenient.

PNT version/Rating strategy	*Mean*	*SD*
Internet-based/Primary strict	72.05	17.33
Internet-based/Secondary strict	83.39	14.39
Internet-based/rimary lenient	74.82	17.82
Internet-based/Secondary lenient	86.06	14.89
Paper-based/Primary strict	78.05	17.29
Paper-based/Secondary strict	85.54	16.07
Paper-based/Primary lenient	83.30	18.40
Paper-based secondary lenient	91.72	17.71
Oral primary	89.18	9.18
Oral secondary	95.66	8.66

[5]Strict condition in rating of the internet- and paper-based PNTs cannot be compared with the oral PNT, because the latter involves no spelling.

Table 2.
Pearson correlations between internet-based, paper-based, and oral PNTs scores obtained by applying four rating strategies: primary strict, secondary strict, primary lenient and secondary lenient.

	Internet-based			Paper-based				Oral	
	2	3	4	5	6	7	8	9	10
1. Internet-based primary strict	.96	1.00	.96	.76	.71	.76	.69	.62	.60
2. Internet-based secondary strict		.95	.99	.68	.66	.67	.63	.61	.58
3. Internet-based primary lenient			.96	.76	.72	.77	.70	.63	.61
4. Internet-based secondary lenient				.67	.66	.67	.64	.61	.58
5. Paper-based primary strict					.97	.99	.96	.68	.68
6. Paper-based secondary strict						.96	.99	.66	.67
7. Paper-based primary lenient							.97	.67	.68
8. Paper-based secondary lenient								.65	.66
9. Oral primary									.98
10. Oral secondary									

All *p*s < .001

were significant predictors of the total score ($F(3, 83) = 57.44$, $p < .001$, adjusted-$R^2 = .67$; $b = 6.39$, $SE = .63$, ß = .69, $t = 10.18$, $p < .001$ for listening; and $b = 1.26$, $SE = .343$, ß = .25, $t = 3.66$, $p < .001$ for reading).

DIALANG

The DIALANG six proficiency levels were ranked 1 through 6 with a greater rank representing higher proficiency level. First, it was found that the correlations between the self-assessment and testing scores in all three modules for which the self-assessment was administered were highly significant (ρ = .32, $p < .01$, for listening; ρ = .43, $p < .001$, for writing; and ρ = .27, $p < .05$, for reading). The correlations between the VSPT and all testing scores were also highly significant (see **Table 4**).

Proficiency Tests Comparison

The iPNT, PPVT, Cloze test, total TOEFL, and DIALANG placement scores significantly correlated with each other (see **Table 5**). The iPNT was also found to significantly correlate with DIALANG scores for listening (ρ = .53, $p < .001$), writing (ρ = .55, $p < .001$), reading (ρ = .57, $p < .001$), structure (ρ = .48, $p < .001$), and vocabulary (ρ = .57, $p < .001$) modules. In addition, it correlated significantly with TOEFL scores for listening ($r = .54$, $p < .001$) and reading ($r = .31$, $p < .01$) modules. The respective TOEFL and DIALANG scores for listening (ρ = .57, $p < .001$), writing (ρ = .54, $p < .001$), and reading (ρ = .41, $p < .001$) also significantly correlated with each other.

Table 3.
Pearson correlations between TOEFL scores.

	2	3	4
1. Listening	.36**	.18	.79***
2. Reading		.27*	.51***
3. Writing			.23*
4. Total			

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 4.
Spearman correlations between DIALANG VSPT and testing scores.

	2	3	4	5	6
1. VSPT	.47	.45	.42	.44	.55
2. Listening		.57	.63	.52	.48
3. Writing			.61	.60	.53
4. Reading				.55	.49
5. Structure					.62
6. Vocabulary					

All *p*s < .001.

Experiment 2

A different group of participants from the same subject pool was recruited for this experiment. The participants were 130 students (45 male and 85 female; aged between 17 and 26, $M = 19.94$, $SD = 1.82$). They were administered the iPNT twice with a 35 days lag between the sessions. The responses were rated using the primary list of labels and strict spelling (see above). A highly significant correlation ($r = .83$, $p < .001$) between the iPNT scores on the first and the second sessions suggests a high test-retest reliability of the assessment tool.

Discussion

The study presents a new psychometric tool assessing language proficiency with respect to an individual's productive vocabulary. The iPNT is an internet-based test that assesses vocabulary knowledge by rating participants' responses to 120 colored drawings of simple objects. Participants are given eight minutes to type the names of the objects, which consequently are compared against a list of correct labels.

To employ an automatic rating of the iPNT, only those responses that perfectly match corresponding items from a list of correct responses were scored a point. To ensure convergent

Table 5.
Pearson correlations between various language proficiency tests.

	2	3	4	5
1. iPNT	.60	.53	.68	.52
2. PPVT		.44	.57	.43
3. Cloze			.60	.47
4. TOEFL total				.45
5. DIALANG VSPT				

All *p*s < .001.

validity of this scoring, four rating strategies were applied to participants' responses: primary strict, primary lenient, secondary strict and secondary lenient. The findings that all four rating strategies provided highly correlated results justify the iPNT rating based on the primary list of labels and strict spelling. According to this rating schema, only those responses that perfectly match correct labels should score a point. Therefore, a simple algorithm implemented in a computer can process the responses and automatically provide a language proficiency score.

Three versions of the PNT—internet-based, paper-based, and oral—were administered to participants. Although, all three versions obtained significantly different scores, they were found to correlate with each other highly. These findings eliminate potential bias that may have occurred due to change in media used to administer the test. The iPNT can be safely used to control for language proficiency in a sample assessed with the same version of the test. However, it is not recommended to use different PNT versions in the same sample.

To test the concurrent validity of the productive vocabulary iPNT as a test of language proficiency, it was compared with the tests assessing different linguistic skills: receptive vocabulary PPVT, overall language proficiency Cloze test, admission TOEFL, and diagnostic system DIALANG. Although correlations between the iPNT and other tests were highly significant, the correlation values ranging from .52 to .68 indicate that this test provides a partial assessment of the linguistic abilities measured by other tests in this study. This limitation is common to most abbreviated language proficiency tests, which is compensated by their efficient administration. Note that the purpose of the iPNT is to assess an individual's productive vocabulary knowledge. Therefore, it can be only used in research that taps into this specific linguistic ability. For example, Kharkhurin (2011) used paper-based PNT to assess bilinguals' vocabulary knowledge and relate it to their performance on the tests of selective attention, creativity, and fluid intelligence. Greater vocabulary knowledge was hypothesized to facilitate certain cognitive mechanisms underlying performance on these tests. In this framework, productive vocabulary test was an appropriate measure of bilinguals' language proficiency.

Another potential limitation of this test stems from its procedure. The participants have to type their responses within a limited time interval, which presents a potential disadvantage for poor typists. The sample of the present study comprised of college students with presumably extensive typing experience, but even in this sample the scores on the paper version were higher than on the internet version when participants' responses were rated using the primary list of labels with strict spelling. Future studies should look into this issue by establishing age and education related norms, which reflect participants' typing abilities.

In conclusion, the current preliminary version of the iPNT presents a new reliable tool that offers a number of advantages to the empirical investigation that involves language proficiency assessment. This test can be administered online in a relatively short period of time (eight minutes) without involvement of any additional resources. The data file with participants' responses can be uploaded into statistical software for further processing. In a new version of this test[6], the iPNT score is calculated within the testing environment and test users are provided with an outcome immediately upon completion of the test. Another important advantage of this test is its suitability for any language. The iPNT can be used in any language providing a list of correct labels in that language. The current version of the test includes an interface option to upload a list of labels in a given language. The administrating convenience and online accessibility of the test encourages future studies to provide the norms for this test by collecting data in a broad range of linguistic and cultural groups.

[6]http://www.harhur.com/research/ipnt.html

REFERENCES

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22,* 301-320. doi:10.1191/0265532205lt310oa

Baldauf, R. B., & Propst, I. K. (1979). Matching and multiple-choice cloze tests. *Journal of Educational Research, 72,* 321-326.

Briere, E. J., Clausing, G., Senko, D., & Purcell, E. (1978). A look at cloze testing across languages and levels. *Modern Language Journal, 62,* 23-26.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977-990. doi:10.3758/BRM.41.4.977

Carrow-Woolfolk, E. (1999). *Comprehensive assessment of spoken language*. Circle Pines, MN: American Guidance Service.

Chapelle, C. A. (2006). DIALANG: A diagnostic language test in 14 European languages. *Language Testing, 23,* 544-550. doi:10.1191/0265532206lt341xx

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge, MA: Cambridge University Press.

Darnell, D. K. (1968). *The development of an English language proficiency test of foreign students, using a clozentropy procedure* (No. Bureau No. BR-7-H-OlO). Boulder, CO: Colorado University.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary Test-IV*. Circle Pines, MN: American Guidance Service.

Dupuis, M. M. (1980). The cloze procedure as a predictor of comprehension in literature. *Journal of Educational Research, 74,* 27-33.

ETS. (2005). *TOEFL® internet-based test: Score comparison tables*. Princeton, NJ: Educational Testing Service.

ETS. (2008a). *Reliability and comparability of TOEFL® iBT scores*. Princeton, NJ: Educational Testing Service.

ETS. (2008b). *TOEFL iBT and PBT: A comparison*. Princeton, NJ: Educational Testing Service.

Jochems, W., & Montens, F. (1987). De multiple-choice cloze-toets als algemene taalvaardigheidstoets. *Tijdschrift voor Onderwijsresearch, 12,* 133-143.

Jongsma, E. A. (1980). *Cloze instruction research: A second look*. Newark, DE: International Reading Association.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston naming test*. Philadelphia, PA: Lea & Febiger.

Kharkhurin, A. V. (2005). *On the possible relationships between bilingualism, biculturalism and creativity: A cognitive perspective.* Unpublished Dissertation, New York: City University of New York.

Kharkhurin, A. V. (2008). The effect of linguistic proficiency, age of second language acquisition, and length of exposure to a new cultural environment on bilinguals' divergent thinking. *Bilingualism: Language and Cognition, 11,* 225-243.

Kharkhurin, A. V. (2011). The role of selective attention in bilingual creativity. *Creativity Research Journal, 23,* 239-254. doi:10.1080/10400419.2011.595979

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lemmon, C. R., & Goggin, J. P. (1989). The measurement of bilingualism and its relationship to cognitive ability. *Applied Psycholinguistics, 10,* 133-155. doi:10.1017/S0142716400008493

Mariën, P., Mampaey, E., Vervaet, A., Saerens, J., & De Deyn, P. P. (1998). Normative data for the Boston naming test in native Dutch-speaking Belgian elderly. *Brain and Language, 65,* 447-467. doi:10.1006/brln.1998.2000

Oller, J. W. Jr. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56,* 151-158. doi:10.2307/324037

Oller, J. W. Jr. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23,* 105-118. doi:10.1111/j.1467-1770.1973.tb00100.x

Oller, J. W. Jr., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21,* 183-195. doi:10.1111/j.1467-1770.1971.tb00057.x

Padilla, A. M., & Ruiz, R. A. (1973). *Latino mental health: A review of literature*. Washington DC: US Government Printing Office.

Patricacou, A., Psallida, E., Pring, T., & Dipper, L. (2007). The Boston naming test in Greek: Normative data and the effects of age and education on naming. *Aphasiology, 21,* 1157-1170. doi:10.1080/02687030600670643

Peterson, J., Peters, N., & Paradis, E. (1972). Validation of the cloze procedure as a measure of readability with high school, trade school, and college populations. In F. B. Greene (Ed.), *Investigations relating to mature readers, twenty-first yearbook of the National Reading Conference* (pp. 45-50). Milwaukee: The National Reading Conference, Inc.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception, 33,* 217-236. doi:10.1068/p5117

Semel, E. M., Wiig, E. H., & Secord, W. (2003). *Clinical evaluation of language fundamentals* (3rd ed.). San Antonio, TX: Psychological Corporation.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6,* 174-215. doi:10.1037/0278-7393.6.2.174

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30,* 415-433.

Williams, K. T. (2001). *Group reading assessment and diagnostic evaluation*. Circle Pines, MN: American Guidance Service.

Williams, K. T. (2007). *Expressive vocabulary test* (2nd ed.). Circle Pines, MN: American Guidance Service.

Zhang, Y. (2008). *Repeater analyses for TOEFL® iBT* (ETS Research Memorandum No. RM.08-05). Princeton, NJ: Educational Testing Service.

Appendix A. Picture Naming Test stimuli

1		[I] rolling pin	23		[I] zebra
2		[I] pen	24		[I] basket
3		[I] umbrella	25		[I] cake
4		[I] nose	26		[I] truck
5		[I] doorknob	27		[I] blouse [II] Shirt [II] jacket
6		[I] box	28		[I] dress
7		[I] bicycle	29		[I] key
8		[I] rabbit	30		[I] nail
9		[I] refrigerator	31		[I] butterfly
10		[I] duck	32		[I] mouse
11		[I] leaf	33		[I] kangaroo
12		[I] coat	34		[I] mountain
13		[I] frog	35		[I] mushroom
14		[I] doll [II] baby [II] little girl	36		[I] hanger
15		[I] screwdriver	37		[I] lamp
16		[I] kettle [II] tea kettle [II] teapot	38		[I] cigar
17		[I] cap	39		[I] balloon
18		[I] pants	40		[I] baby carriage [II] carriage
19		[I] brush	41		[I] chair
20		[I] sweater	42		[I] eye
21		[I] pineapple	43		[I] dresser [II] bureau [II] chest [II] chest of drawers
22		[I] snake			
44		[I] pear	68		[I] clown
45		[I] bell	69		[I] watermelon
46		[I] hat	70		[I] anchor
47		[I] grapes	71		[I] rooster [II] chicken
48		[I] fork	72		[I] wine glass [II] glass [II] goblet
49		[I] helicopter	73		[I] chicken [II] hen
50		[I] light bulb	74		[I] pipe
51		[I] ruler	75		[I] frying pan [II] pan

52		[I] seal	76		[I] windmill
53		[I] car [I] Lincoln	77		[I] corn
54		[I] wrench	78		[I] moon [II] quarter moon [II] crescent moon [II] half moon
55		[I] rhinoceros			
56		[I] donkey	79		[I] saltshaker
57		[I] hammer	80		[I] arrow
58		[I] horse	81		[I] turtle
59		[I] whistle	82		[I] harp
60		[I] sandwich	83		[I] stool
61		[I] sock	84		[I] church
62		[I] rocking chair	85		[I] nut
63		[I] hand	86		[I] motorcycle
64		[I] strawberry	87		[I] flower
65		[I] clothespin	88		[I] traffic light [II] stop light
66		[I] paintbrush [II] brush	89		[I] goat
67		[I] flag	90		[I] cup
91		[I] camel	106		[I] spoon
92		[I] train	107		[I] television [II] tv [II] television set
93		[I] ant	108		[I] pencil
94		[I] dog	109		[I] wheel
95		[I] toothbrush	110		[I] iron
96		[I] swan	111		[I] apple
97		[I] saw	112		[I] scissors
98		[I] violin	113		[I] canon
99		[I] spool of thread [II] thread [II] spool	114		[I] shirt
100		[I] baseball bat [II] bat	115		[I] caterpillar
101		[I] star	116		[I] owl
102		[I] cigarette	117		[I] bus
103		[I] pitcher	118		[I] beetle
104		[I] envelope	119		[I] axe
105		[I] drum	120		[I] light switch

[I] indicates the words from the primary list and [II] the words from the secondary list.