Morphological and phonological processing in English monolingual, Chinese-English bilingual, and Spanish-English bilingual children: An fNIRS neuroimaging dataset

Authors
Xin Sun*[1], Kehui Zhang[1], Rebecca Marks[2], Zachary Karas[1], Rachel Eggleston[1], Nia Nickerson[1], Chi-Lin Yu[1], Neelima Wagley[3], Xiaosu Hu[1], Valeria Caruso[1], Tai-Li Chou[4], Teresa Satterfield[1], Twila Tardif[1], Ioulia Kovelman[1]

Affiliations
[1] Department of Psychology, University of Michigan Ann Arbor
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[3] Department of Psychology and Human Development, Vanderbilt University
[4] Department of Psychology, National Taiwan University

Corresponding author's email address and Twitter handle
Email address: sunxin@umich.edu
Twitter handle: @xinsun1007

Keywords
fNIRS, children, bilingualism, morphological awareness, phonological awareness, reading

Abstract
This article documents a functional Near-Infrared Spectroscopy (fNIRS) neuroimaging dataset deposited in Deep Blue Data. The dataset included neuroimaging and behavioral data from *N* = 343 children aged 5-10 with a diverse linguistic background, including children who are English monolingual, Chinese-English, and Spanish-English bilingual. Children completed phonological and morphological awareness tasks in each of their languages during fNIRS neuroimaging. They also completed a wide range of language and reading tasks. Parents filled in questionnaires to report children's demographic information as well as their home language and literacy backgrounds. The dataset is valuable for researchers in the field of developmental cognitive neuroscience to further investigate questions such as the effects of bilingualism on children's neural basis for literacy development.

Specifications Table

Subject	Developmental and Educational Neuroscience
Specific subject area	fNIRS neuroimaging of morphological and phonological awareness in English monolingual, Chinese-English, and Spanish-English bilingual children
Type of data	Tables, Images, fNIRS hemodynamic data
How data were acquired	TechEn Inc. CW6 fNIRS system with 690 and 830 nm wavelengths, 12 signals, 24 detectors, 46 channels. E-Prime software was used to display stimuli and collect data.
Data format	fNIRS data with block stimuli marks are stored in .nirs files; Proficiency/demographic raw data are stored in excel sheets.
Parameters for data collection	All participants are children growing up in the US and attending English-only schools. The monolingual participants are all native speakers of English and only speak English. The bilingual participants have Spanish or Chinese exposure from home since birth.
Description of data collection	Participants (*N* = 343) completed a behavioral session and a neuroimaging session. The behavioral session assessed participants' language and reading proficiency in each of their languages. The neuroimaging session asked participants to complete morphological and phonological awareness tasks in each of their languages during fNIRS scanning.
Data source location	University of Michigan, Department of Psychology, Ann Arbor, MI.
Data accessibility	Repository: Deep Blue Data Link to data Persistent Identifier: https://doi.org/10.7302/kxgf-ps11
Related research article	1. Sun, X., Zhang K., Marks, R., Nickerson, N., Eggleston, R., Yu, C.L., Chou., T., Tardif, T., & Kovelman, I. (2021). What's in a word? Cross-linguistic influences on Spanish-English and Chinese-English bilingual children's word reading development. *Child Development 93*(1), 84-100. http://doi.org/10.1111/cdev.13666 This article used data from the behavioral assessments of *N* = 283 participants from the current dataset. 2. Sun, X., Marks, R., Zhang, K., Yu, C.L., Eggleston, R., Nickerson, N., Chou, T.L., Hu, X.S., Tardif, T., Satterfield, T., & Kovelman, I. (In Press). Brain bases of English morphological processing: A comparison between Chinese-English, Spanish-English bilingual, and English monolingual children. *Developmental Science* Preprint available at: https://osf.io/9zx2t/ 3. Marks, R. A., Eggleston, R., Sun, X., Yu, C. L., Zhang, K., Nickerson, N., Hu, X., & Kovelman, I. (2021). The neurobiological basis of morphological processing for typical and impaired readers. *Annals of Dyslexia* https://doi.org/10.1007/s11881-021-00239-9

	This article used data of the fNIRS English morphological awareness task as well as the corresponding behavioral data of N = 97 English monolingual participants from the current dataset.

Value of the Data (400 characters per point)

- Bilingualism research will benefit from this developmental dataset of young Spanish-English and Chinese-English bilinguals, allowing for inquiries into the effects of age of acquisition, experience, proficiency, and cross-linguistic transfer in children's emerging neural architectures for language and literacy development.
- The dataset will equip researchers in the fields of developmental, educational, and cognitive neuroscience to address questions about children's neuro-cognitive profiles for language and literacy development across three typologically-distinct languages.
- The dataset is extensive and allows for investigations into (but not limited to) meaningful research topics: the neural basis of phonological and morphological skills, behavioral indicators associated with the developing language brain networks, and the neural and behavioral profiles of children from diverse backgrounds such as those with bilingual experiences, dyslexia or reading disabilities.

1. Data Description

All data (raw neuroimaging data, neuroimaging task accuracy and reaction time, behavioral assessment raw and standard scores, and demographics) are available in the DeepBlue repository under the name "Morphological and phonological processing in English monolingual, Chinese-English bilingual, and Spanish-English bilingual children: An fNIRS neuroimaging dataset". For a list of the Deep Blue files and contents, see Table 1.

Table 1. Full list of the Deep Blue Data files and contents

Data/Measure	File Name in Deep Blue	Data/Measure Content
fNIRS imaging	Chinese_NIRSfiles.zip	.nirs files by ID and task for Chinese-English bilinguals
	English_NIRSfiles.zip	.nirs files by ID and task for English monolinguals
	Spanish_NIRSfiles.zip	.nirs files by ID and task for Spanish-English bilinguals
	NIRSfile_Readin_Plot.m	A Matlab script that helps import and plot .nirs files into the Matlab program
Task performance	Task_Performance_Data.zip	Excel spreadsheets including behavioral task performance (1 file), fNIRS task accuracy (2 files) and reaction time (2 files)

Demographics	Participant_Demographics.xlsx	Demographic information, including age of testing, gender, grade, etc.
Language and literacy backgrounds	Language_and_Literacy_ Background (ILQ, BOQ).xlsx In-Lab_Questionnaire_ILQ.pdf Bilingual_Outcomes_ Questionnaire_(BOQ)_English Spanish. pdf Bilingual_Outcomes_Questionnaire _(BOQ)_Chinese. pdf	Itemized data for the In-Lab Questionnaire and the Bilingual Outcomes Questionnaire. Full In-Lab Questionnaire (ILQ) Full Bilingual Outcomes Questionnaire (BOQ) in English, Spanish, and Chinese
Behavioral measures	Self-developed_Behavioral _Measures.zip	All self-developed behavioral measure items

Neuroimaging data are raw data files with block stimuli marks that signify on-task periods task condition. The neuroimaging data folder was organized by participant group and task. Specifically, under the folder “NIRS files”, subfolder “Chinese” includes all fNIRS data for the Chinese-English bilingual children, subfolder “English” is for the English monolingual children, and subfolder “Spanish” is for the Spanish-English bilingual children. There are two folders in the “English” subfolder, and four folders in the “Chinese” and “Spanish” subfolders that include data for specific tasks. For example, folder “English Morphology” includes the fNIRS data for the English morphological awareness task, folder “Chinese Phonology” includes the fNIRS data for the Chinese phonological awareness task. Under these folders, each fNIRS file is stored in an individual folder named after participant ID. For example, file “3007_CH_MA.nirs” is stored in folders “NIRS files” – “Chinese” – “Chinese Morphology” – “3007” and it is the fNIRS file for participant 3007 during their Chinese morphological task. All fNIRS neuroimaging data are *.nirs* files and can be easily read into most Matlab scripts. Table 1 shows the number of participants who completed each neuroimaging task by language group.

The “Task Performance Data.zip” includes all behavioral performance for the neuroimaging and behavioral assessments, presented with excel sheets. Neuroimaging task accuracy and reaction time are presented in two Excel sheets, named “R01_E-Prime Accuracy.xlsx” and “R01_E-Prime Reaction Times.xlsx”, respectively. The neuroimaging task items are included in the sheets (see the “read me” sheet in the excel files). Raw and standard scores for the behavioral assessments are also provided in an Excel sheet named “R01_Behavioral Measures.xlsx”. All self-developed behavioral assessments are presented in “Self-developed Behavioral Measures.zip”.

Demographic and language background data are presented in two Excel sheets, named “Participant_Demographics.xlsx” and “Language_and Literacy_Background(ILQ, BOQ Data).xlsx”. The latter data sheet includes data from two questionnaires, and the full list of questionnaire items are presented with two word documents, named “In-Lab Questionnare (IBQ).docx” and “Bilingual Outcomes Questionnaire(BOQ).doc”.

Table 2. Number of Participants by fNIRS Neuroimaging Task by Language Group

Task	Number of Participants (*N*)		
	Monolingual	Chinese Bilingual	Spanish Bilingual

English Morphological Awareness	131	99	104
English Phonological Awareness	114	98	96
Chinese Morphological Awareness	/	94	/
Chinese Phonological Awareness	/	89	/
Spanish Morphological Awareness	/	/	96
Spanish Phonological Awareness	/	/	93

Note. This table displays the number of participants in the fNIRS task. The numbers mostly but not fully align with the behavioral task.

2. Experimental Design, Materials, and Methods

2.1 Participants

Participants included *N* = 343 children aged 5 to 11 (M_{age} = 8.08, SD_{age} = 1.64, 161 girls). Participants were divided into three groups according to their language experience. All monolinguals were born to native English speakers and exposed to English-only language environments. Bilingual participants had at least one parent as a native speaker of either Chinese or Spanish and were exposed to the language at home, from birth. The English monolingual group included *N* = 135 children aged 5.4 to 11.9 (M_{age} = 8.46, SD_{age} = 1.65, 64 girls); the Chinese-English bilingual group included *N* = 102 children aged 5.1 to 11.5 (M_{age} = 7.51, SD_{age} = 1.67, 46 girls); and the Spanish-English bilingual group included *N* = 106 children aged 5.7 to 11 (M_{age} = 8.13 , SD_{age} = 1.44, 51 girls). Within the English monolingual group, *N* = 8 were delayed in reading (M_{age} = 9.22, SD_{age} = 1.16, 2 girls), as indicated by their standard scores below 85 in at least two of the four reading tasks (i.e., Word Reading. Word Attack, Reading Comprehension, and Reading Fluency; and *N* = 20 had dyslexia (M_{age} = 9.45; SD_{age} = 1.61, 11 girls), as indicated by their 1) standard scores below 85 in at least two reading tasks, and 2) PPVT standard score 2 standard deviations (30 points) higher than word reading.

2.2 Behavioral assessments and the demographic information

Participants completed behavioral assessments in each of their languages while their parents filled out demographic questionnaires. The behavioral testing assessed key language and literacy skills including phonological awareness, morphological awareness, vocabulary, single-word reading, nonword reading, passage comprehension, and sentence reading fluency. The format of the heritage language measures maximally paralleled the English tasks. In addition, a backward digit span task was administered in English (WISC-V, Wechsler, 2014[1]). Details of language and literacy measures are shown in Table 2. All self-developed measures can be found in the data repository.

Table 2. Language and literacy measures by language

	English		Spanish		Chinese	
Construct	Measure	Reference	Measure	Reference	Measure	Reference
Phonological awareness	Comprehensive Test of Phonological Processing Elision	Wagner et al. (1999)[2]	Test of Phonological Processing in Spanish (TOPPS)	Francis et al. (2001)[5]	Self-developed Syllable and Phoneme Elision task	Newman et al. (2011)[3]; Sun et al. (2021)[4]
Morphological awareness	Self-developed Early Lexical Morphology Measure (ELMM)	Adapted from Goodwin et al. (2012)[6]	Self-developed Early Lexical Morphology Measure -Spanish (ELMM-S)	Modeled after the English task	Self-developed Morphological Construction Test	Song et al. (2015)[7]; Sun et al. (2021)[4]
Vocabulary	Peabody Picture Vocabulary Test-5 (PPVT)	Dunn (2015) [8]	Test de Vocabulario en Imágenes Peabody (TVIP)	Dunn et al. (1986)[10]	Peabody Picture Vocabulary Test-Revised	Lu & Liu (1998)[9]
Nonword reading	Woodcock Johnson-4 Word Attack Subset (WJ-WA)		/	/	/	/
Single-word reading	Woodcock Johnson-4 Letter-word Identification Subset (WJ-LWID)		Batería III Woodcock-Muñoz Identificacion de letras y palabras		Self-developed Character Recognition and Reading Task	Sun et al. (2021)[4]
Passage comprehension	Woodcock Johnson-4 Passage Comprehension Subset (WJ-PC)	Schrank et al., 2018 [11]	Batería III Woodcock-Muñoz Comprehension de textos	Muñoz-Sandoval et al. (2005) [12]	/	/
Sentence reading fluency	Woodcock Johnson-4 Sentence Reading Fluency Subset (WJ-SRF)		Batería III Woodcock-Muñoz Fluidez en la lectura		Self-developed Sentence Reading Fluency Task	/

2.3 fNIRS imaging tasks

Participants completed a morphological awareness and a phonological awareness task in each of their languages during fNIRS scanning. All of the tasks followed a block design and each lasted 7.2 minutes. Each task had twelve 30-second blocks and each block displayed 4 items, yielding 48 items in total. Blocks were separated by a 6-second break. All of the tasks had 3 conditions: 2 experimental conditions and 1 control condition. Each condition had 4 blocks (16 items). Blocks were presented with a fixed sequence and blocks of the same condition were not presented in succession. All task items followed the same paradigm: First, participants heard three words; next, they were asked to select which word of the last two matched the first (target) word by pressing a button. To help participants focus on the words they heard, the computer screen presented a colored box in place of the word stimulus (See Figure 1). All tasks were presented with E-Prime.

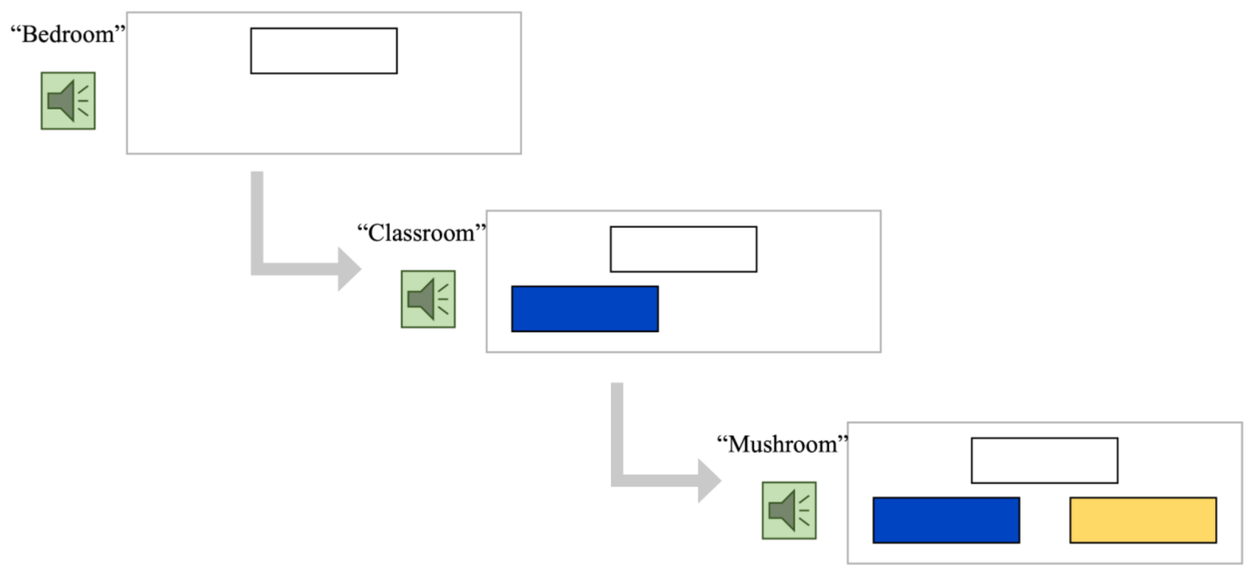

Figure 1. Sample screen display of an English Morphological awareness item
Note. Participants would see a blank box display as they heard each word. The top box corresponded to the target word while the bottom two boxes corresponded to the two words of choice.

2.3.1. Morphological awareness task. The morphological awareness task asked participants to select the word that matched the meaning of the target word. For each item in the experimental conditions, the correct answer shared a morpheme with the target word while the distractor had a syllable that sounded identical but did not share a meaningful component with the target word. Experimental condition 1 was a *compound* condition that consisted of compound word targets. An English example is *classroom*, *bedroom*, *mushroom*; a Chinese example is *朋友* (/peng2 you3/ friend)， *好友* (/hao3 you3/ good friend)， *没有* (/mei2 you3/ none); a Spanish example is *mar* (sea), *marinero* (sailor), *mariposa* (butterfly). Experimental condition 2 was a *derivational* condition that presented derivational word targets. An English example is *runner*, *juggler*, *flower*; a Chinese example is *读者* (/du2 zhe3/ reader)， *记者* (/ji4 zhe3/ journalist)， *或者*(/huo4 zhe3/ or); a Spanish example is *expresidente* (expresident), *exnovio* (ex-boyfriend), *examen* (test). The control condition was a word recognition task. For each item, one of the last two words would be identical to the target word. For example, *number*, *number*, *taxi*. The full list of items can be found in the Excel sheets for the neuroimaging task accuracy and reaction time.

2.3.2 Phonological awareness task. The phonological awareness tasks asked participants to select the word that matched the first sound of the target word. For each item in the experimental conditions, the correct answer would share the first sound with the target word, while the distractor would be semantically related but shared no initial sound with the target word. Experimental condition 1 was the *easy* condition. Words in this condition were less difficult: they did not have glides or diphthongs (in English and Spanish), and/or the distractor initial sounds were phonetically distant from the target words. An English example is *mother*, *major*, *father*; a Chinese example is *半夜* (/ban4 ye4/ midnight), *毕业* (/bi4 ye4/ graduate), *深夜* (/shen1 ye4/ late night); a Spanish example is *salmón* (salmon), *camarón* (shrimp), *pantalón* (pants). Experimental condition 2 was the *hard* condition. Words in this condition were more difficult: they had either glides or diphthongs and/or the distractor initial sounds were phonetically similar to the target words. An English example with glide is *teeth*, *truth*, *mouth*; a Chinese example with a harder distractor is *帽子* (/mao4 zi/ midnight), *面子* (/mian4 zi/ face/), *脑子* (/nao3 zi/ brain); a Spanish example is *lunes* (Monday), *leones* (lions), *jueves* (Thursday). The control condition was identical to that in the morphological awareness task, but with different words. The full list of items can be found in the Excel sheets for the neuroimaging task accuracy and reaction time.

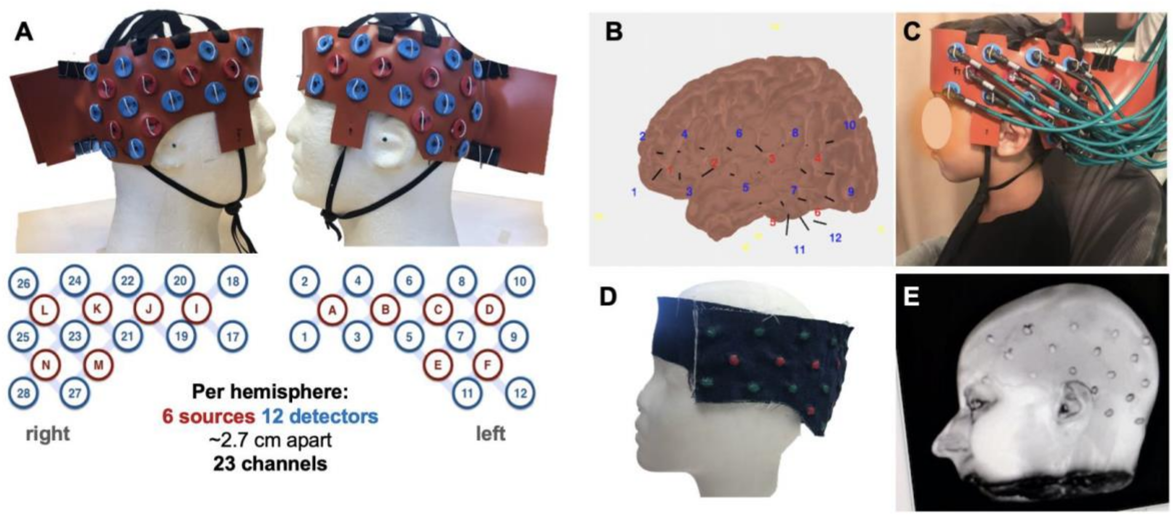

Figure 2. fNIRS cap configuration. (A) how signal (red, letters) and detector (blue, numbers) sensors are located on a silicone-rubber band around the participant's head, (B) map of the estimated brain regions covered by the cap design as digitized using AtlasViewer GUI (Aasted et al., 2015), (C) participant wearing the cap during data acquisition, (D) MRI version of the cap with vitamin-e capsules, and (E) visualization of vitamin-e capsules on the skull.

2.4 fNIRS data acquisition

The fNIRS cap set-up included 12 emitters of near-infrared light sources and 24 detectors spaced ~2.7 cm apart, yielding 46 data channels (i.e., source-detector pairings; 23 channels per hemisphere; see Figure 2). The light sources and detectors were mounted onto a custom-built head cap constructed from 2 mm silicone rubber material with grommet attachments. The source and detector alignments were placed precisely in a grid-like formation, ensuring full coverage of the participant's frontal, temporal, and temporoparietal regions across multiple channels. The probes were applied as uniformly as possible for every participant using the international 10-10 transcranial system positioning (Jurcak, Tsuzuki, & Dan, 2007[13]); nasion, inon, Fpz, and left and right pre-auricular points, head circumference were measured and F7, F8, T3, and T4 were anchored to a specific source or detector. Once all optodes were placed on the cap, digital photos of the participant's head and cap alignment were taken from the left, right, and center midline angles.

TechEn-CW6 software signal-to-noise ratio (SNR) minimum and maximum were set to the standard 80 dB and 120 dB power range, respectively. Before the start of each experimental task, the data quality control check was completed by detecting the participant's cardiac signal across key channels of interest and ensuring the fNIRS signals were within the power parameters. When required, the experimenters would adjust the positioning of the cap or participant's hair to register an apt cardiac signal. Data were collected at a sampling frequency of 50Hz.

Ethics Statements

Informed consent was obtained from all participating children and their guardians. In addition, all research protocols were approved by the Institutional Review Board at the University of Michigan Ann Arbor and the protocol number is HUM00033727. The dataset has also removed all identifiable information to protect participant privacy.

CRediT author statement

Xin Sun: Measure development, Data curation, validation, writing - original draft;
Kehui Zhang and **Rebecca Marks**: Measure development, Data curation, validation, writing - review and editing;
Ioulia Kovelman: Conceptualization, Methodology, Supervision, Funding acquisition, writing - review & editing;
All others: Data curation, validation, writing - review and editing.

Acknowledgments

The authors thank members of the Language and Literacy Laboratory at the University of Michigan who helped with participant recruitment, scheduling, and data acquisition. We also thank the National Institutes of Health for funding this work (Kovelman, PI: R01HD092498).

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D.Wechsler, Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V), The Psychological Corporation, 2014.
[2] R. K.Wagner, J. K.Torgesen, C. A. Rashotte, & N. A. Pearson, Comprehensive test of phonological processing: CTOPP, Pro-ed, 1999.
[3] E. H. Newman, T. Tardif, J. Huang, & H. Shu, Phonemes matter: The role of phoneme-level awareness in emergent Chinese readers, Journal of Experimental Child Psychology. 108 (2011) 242–259. https://doi.org/10.1016/j.jecp.2010.09.001
[4] X. Sun, K. Zhang, R. A. Marks, N. Nickerson, R.L. Eggleston, C.L. Yu, T. L. Chou, T. Tardif, & I. Kovelman, What's in a word? Cross-linguistic influences on Spanish– English and Chinese–English bilingual children's word reading development, Child Development, 93 (2021) 84-100. https://doi.org/10.1111/cdev.13666
[5] D. Francis, M. Carlo, D. August, D.Kenyon, V. Malabonga, S. Caglarcan, & M. Louguit, Test of Phonological Processing in Spanish, Center for Applied Linguistics, 2001.
[6] A. P. Goodwin, A. C. Huggins, M. Carlo, V. Malabonga, D. Kenyon, M. Louguit, & D. August, Development and validation of extract the base: An English derivational morphology test for third through fifth grade monolingual students and Spanish- speaking English language learners, Language Testing, 29 (2012) 265–289. https://doi.org/10.1177/0265532211419827
[7] S. Song, M. Su, C. Kang, H. Liu, Y. Zhang, C. McBride-Chang, T. Tardif, H. Li, W. Liang, Z. Zhang, & H. Shu, Tracing children's vocabulary development from preschool through the school-age years: An 8-year longitudinal study, Developmental Science, 18 (2015) 119–131. https://doi.org/10.1111/desc.12190
[8] D.M. Dunn, Peabody Picture Vocabulary Test 5, NCS Pearson, 2015.
[9] L. Lu, & H. S. Liu, The Peabody Picture Vocabulary Test– revised in Chinese, Psychological Publishing, 1998.
[10] L. Dunn, F. Padilla, D. Lugo, & L. Dunn, TVIP: Test Vocabolario Imágenes Peabody, American Guidance Service, 1986.
[11] F. A. Schrank, K. S. McGrew, & N. Mather, Woodcock-Johnson IV, Riverside, 2014.
[12] A. F. Muñoz-Sandoval, R.W. Woodcock, K. S. McGrew, & N. Mather, Batería III Woodcock-Muñoz, Riverside Publishing, 2005.
[13] V. Jurcak, D. Tsuzuki, & I. Dan, 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems, Neuroimage, 34 (2007) 1600-1611. https://doi.org/10.1016/j.neuroimage.2006.09.024