

AMERICAN UNIVERSITY OF BEIRUT

DEVELOPING THE FIRST LEBANESE PICTURE NAMING
TEST: A PRELIMINARY STUDY

by

RAMA RAND NADIM KANJ

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Arts
to the Department of Education
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
October 2018

AMERICAN UNIVERSITY OF BEIRUT

DEVELOPING THE FIRST LEBANESE PICTURE NAMING
TEST: A PRELIMINARY STUDY

by

RAMA RAND NADIM KANJ

Approved by:

II. Thesis or Project Committee (professorial rank only)

Thesis or Project Advisor: Dr. Karma el-Hassan

Rank: Associate Professor

Signature:



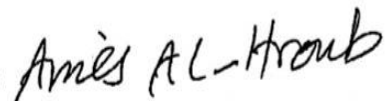
Member: Dr. Ra Zeinoun
Rank: Assistant Professor

Signature:



Member: Dr. Amies Al-Hraib
Rank: Associate Professor

Signature:



Member: Dr. Tariq Khwaileh
Rank: Assistant Professor

Signature:



Date of thesis defense: October 17, 2018

ACKNOWLEDGMENTS

Working on this project has been challenging and extremely rewarding. I would like to thank a number of people who have contributed to the study's outcome in many different ways. I would first like to thank my program and thesis advisor Dr. Karma El-Hassan for her continuous support throughout the past three years, immense knowledge in the field of tests and measurements, readiness to meet and discuss new findings at any time, and trust in my project from the very first day.

I would like to thank the thesis committee members: Dr. Anies el-Hroub, chair of the Department of Education at AUB, for accepting to join the committee despite a full and overflowing schedule, and Dr. Tariq Khwaileh for his valuable time and insightful comments all the way from Qatar University. My special thanks goes to Dr. Pia Zeinoun for nurturing my enthusiasm in the field of assessments, willingness to share her expertise, giving me the freedom to complete my thesis alongside my full-time job and actively listening to my endless worries and concerns.

I would also like to thank the experts who were involved in the development of the psycholinguistic database: Ms. Hala Raad, Ms. Dana Abdul-Ahad, Ms. Doha Berjawi, Dr. Rola Balaa, Ms. Ruba Abou El-Hosn, Ms. Leila Ghoussoub, Ms. Mayssa Boubess, Ms. Sara Bitar, Ms. Cynthia Roukoz, and Ms. Aya Hajj. Without their generous participation and input, this project would not have been possible.

I thank my fellow colleagues at the Psychological Assessment Center at AUBMC, Ms. Cynthia Roukoz, Dr. Rasha Mashmoushi and Dr. Marc Barakat for their patience and friendly support. I thank my partner Dr. Joseph El-Khoury for his continued encouragement, and willingness to proofread countless pages of psychotechnical terms.

I would like to extend thanks to individuals in many countries, who so generously contributed to the work presented in this thesis. I thank Dr. Dragos Iliescu, Dr. Jon Andoni Duñabeitia and Dr. Sumru Erkut for answering questions through the process of researching and writing this thesis.

I thank my "critical friends", Ms. Cheryl Moawad, Ms. Rouba Khalaf and Ms. Linda Bou Ali, for their constructive feedback and inquisitive questions throughout the process which, without them knowing, incited me to widen my research from various perspectives.

My profound gratitude goes to my parents Dr. Randa Farhat and Dr. Nadim Kanj and siblings Ms. Riwa Kanj and Dr. Amjad Kanj and his wife Dr. Nadine Abdallah for their unlimited support, unconditional care and blind trust in me.

I would like to dedicate this work to children from all corners of the world who, at some point in their lives, endured the consequences of biased assessments. May we allow research to pave the way to a fair world.

AN ABSTRACT OF THE THESIS OF

Rama Rand Nadim Kanj for Master of Arts
Major: Education – Educational Psychology Tests
and Measurements

Title: Developing the First Lebanese Picture Naming Test: A Preliminary Study

Naming ability is a significant predictor of cognitive performance and it plays an essential role in neuropsychological and academic evaluations. At present, there are no standardized tests that are culturally and linguistically appropriate to assess naming ability in lebanese children. Practitioners in lebanon still rely on naming tests developed in western countries, which may threaten the validity of the test results and lead to erroneous conclusions. One solution is to develop a picture-naming test that is culturally suitable for lebanese children.

The purpose of the study is threefold. First, obtain the first lebanese database of psycholinguistic variables for a set of 219 picture words based on eight experts' ratings of cultural familiarity, name agreement, word frequency and age of acquisition. Second, based on the ratings, develop and pilot the first draft of the picture-naming test for typically developing lebanese children between the ages of 3 and 9 years enrolled in private and public schools in beirut. Finally, implement modifications to the test based on test item parameters and results derived from the piloting phase. The test construction method adopts a dual-focus approach for test development in order to develop items in arabic, english and french simultaneously and reduce linguistic and cultural biases. The first-draft picture-naming test was piloted on 74 lebanese children between the ages of 3-0 and 9-11 enrolled in private and public school in beirut. Results were analyzed at the sample level and the item level. Further test revisions are suggested and future directions are outlined.

CONTENTS

ACKNOWLEDGEMENTS	v
ABSTRACT.....	vi
LIST OF ILLUSTRATIONS.....	xii
LIST OF TABLES.....	xiii
Chapter	
I. OVERVIEW OF THE CURRENT STUDY	1
A. Background.....	1
1. Picture Naming.....	2
B. Statement of the Problem.....	3
1. The Risk of Test Bias.....	5
2. Practical Framework for Test Construction: the Dual-Focus Approach.....	6
3. The Need for a Lebanese Psycholinguistic Database.....	7
C. The Current Study.....	9
1. Purpose of the Current Study.....	9
2. Outline of the Methods and Procedure.....	10
3. Participants.....	11
D. Test and Item Analysis	12
1. Analysis at the Sample Level.....	12
2. Analysis at the Item Level.....	12
E. Significance and Need of a Lebanese Picture Naming Test	13
1. Assumptions and Foreseen Limitations.....	14
F. Summarizing the Current Study.....	14
II. LITERATURE REVIEW.....	16
A. Part 1: Language Development, Naming and Lexical Retrieval Models	16

1. First Words.....	16
2. Picture Naming as a Measure of Expressive Vocabulary.....	17
3. An Overview on Lexical Retrieval Models.....	18
B. Part 2: Lebanon, the Case of a “Unique Multilingual and Multicultural Make Up”.....	20
1. Introducing the Lebanese Dialect.....	20
2. Lebanon’s “Two-Faced” Educational System.....	21
3. Implications of Bilingualism on Language Testing.....	22
C. Part 3: Testing Biases and Implications on Assessments of Vocabulary...	25
1. Background on Cross-Cultural Assessment and the Emergence of Test Bias.....	25
2. Bias in Cross-Cultural Assessment: Definition and Types.....	27
3. Current Practices in Translating and Adapting Multicultural Assessment.....	29
4. The Dual-Focus Approach.....	33
5. Statistical Methods to Detect Bias in Assessment.....	34
D. Part 4: Psycholinguistic Databases and Naming Tests in the Arab World	35
1. Describing Psycholinguistic Variables.....	35
2. Psycholinguistic Databases for Arabic Words.....	36
3. Multi-Cultural Psycholinguistic Database.....	37
E. The Ideal Picture Set.....	38
1. The MultiPic Databank.....	39
III. METHODS	42
A. Phase 1: Test Construction	42
1. Selecting the Picture Set for the Present Study.....	42
2. Selecting the Expert Committee Members.....	44
3. Developing the Materials: Rating Booklets.....	45
4. Processing the Data on the Psycholinguistic Variables.....	48
5. Compiling the First Draft.....	49
B. Phase 2: Test Piloting	50
1. Participants.....	51
2. Test Administration and Data Collection.....	52
3. Data Entry and Preprocessing.....	54
C. Data Analysis	55
1. Descriptive Statistical Analysis.....	55

2. Decisions to Retain and Revise Items.....	57
D. Summarizing the Steps of the Methodology.....	59
IV. RESULTS.....	61
A. Outcomes of Test Development.....	61
1. Development of the Psycholinguistic Database for Lebanese Words.....	61
2. The First Draft of the Picture-Naming Test.....	63
B. Outcomes of Test Piloting	67
1. Pilot Sample Characteristics and Descriptive Results.....	61
2. Preliminary Analysis: Exploring Assumptions.....	63
3. Data Analysis at the Sample Level.....	70
a. Comparing Means across Age Group.....	70
b. Comparing Means across Gender and Type of Schooling...	71
c. Predictive Analysis.....	72
4. Data Analysis at the Item Level.....	73
	80
V. DISCUSSION.....	
A. Summarizing the Outcomes of the Study.....	80
1. A Psycholinguistic Database for Lebanese Words.....	80
2. Introducing the Lebanese Picture-Naming Test.....	81
3. Comparison of Test Performance across Group Variables.....	81
4. Analysis at the Item Level.....	84
5. Test Development Procedure and Best Practice Methods.....	84
B. Limitations.....	85
C. Future Directions.....	87
D. Conclusion.....	89

Appendix

I. Appendix A: Rating Booklet Template.....	112
II. Appendix B: Lebanese Psycholinguistic Database for 219 Picture Words.....	114

III.	Appendix C: Decisions to Include, Discard or Examine Items based on Item Parameters after Piloting and Expert’s Comments.....	125
IV.	Appendix D: Decision to Retain or Review Items.....	136
V.	Appendix E: List of the Final Words in the Lebanese Picture Naming Test.....	138
	REFERENCES.....	91

LIST OF ILLUSTRATIONS

Figure		Page
1.	Steps in Creating a Bilingual Measure Using the Dual-Focus Approach (Erkut et al., 1999).	9
2.	Examples of pictures from the MultiPic Database from various semantic categories	40
3.	Flowchart of the Systematic Process for Picture Selection	46
4.	Implementation of the Dual-Focus Approach (Erkut et al., 1999) in the development of the first Lebanese picture-naming test.	55
5.	Figure 5. Decision Tree to Select the Final Items in the LPNT	72

LIST OF TABLES

Table		Page
1	Members of the Committee of Experts.....	42
2	Distribution of the Sample across Age and Gender.....	48
3	List of Discarded Pictures based on Average Cultural Familiarity Rating.....	58
4	List of Discarded Pictures based on percent Name Agreement.....	59
5	Proportions of Semantic Categories.....	61
6	Distribution of the Sample Across Gender and Type of Schooling Separately.....	62
7	Descriptive of the Sample Characteristics Age and Total Score.....	62
8	Mean Performance across Age Groups.....	63
9	Mean Performance stratified by Age Group, Gender and Type of Schooling.....	63
10	Robust Tests of Equality of Means.....	65
11	Multiple Comparisons of Total Score using Games-Howell.....	65
12	Mann-Whitney U test- Gender Differences across Total Score.....	66
13	Mann-Whitney U test- Schooling Differences across Total Score.....	66
14	Regression Parameters.....	67
15	Internal Reliability Measures.....	67
16	Matrix of Test Item Parameters: Item Difficulty Index x Item Discrimination Index (Ages: 3-5).....	69
17	Matrix of Test Item Parameters: Item Difficulty Index x Item Discrimination Index (Age: 6-7).....	70
18	Matrix of Test Item Parameters: Item Difficulty Index x Item Discrimination Index (Ages: 8-9).....	71

Developing the First Lebanese Picture Naming Test:

A Preliminary Study

Chapter 1: Overview of the Current Study

Background

Vocabulary is at the heart of spoken and written language. It is an essential component of effective communication. One way to measure vocabulary is through a task of naming. Naming is a basic human ability that is fundamental for communication through language (Terrace, 1985). In children, it is the earliest step in linguistic production (Etard et al., 2000). Naming is usually measured through a task of *picture naming* (also called *confrontation naming*) where an individual is shown a picture of an object, an action or a concept and is asked to provide the vocabulary word that corresponds to the picture. This task is usually part of a standardized measure that systematically evaluates an individual's performance and allows norm-referenced comparisons. Lexical retrieval models suggest three assumingly sequential cognitive stages that underlie naming: object identification, name activation and response generation (Paivio, Clark, Digdon, & Bons, 1989). Research over the years shows that naming significantly affects other human abilities and predicts cognitive functioning in children and adults. This chapter will first briefly cover important findings on naming in the context of the school and clinical setting. It will then outline the study's aims, methods of test construction, participants, piloting phase and how we plan to analyze results in order to carry out further modifications to the test.

Picture Naming

Research shows that naming plays an important role in areas of childhood development, academic performance and screening for neurodevelopmental disorders and neurocognitive disorders. Naming evolves with age. Studies show that as children develop, they begin to acquire words with increasing length and complexity (Ilkman, 2015; Spinelli et al., 2005). Gender effect on naming has been inconsistent with some studies showing better naming abilities in females and others in males (Grabowski, Damasio, Eichhorn, & Tranel, 2003; Pineda et al., 2000). In the context of academic performance, the relationship between naming and school achievement was well established over 40 years ago. An early study by Jansky and Hirsch (1972) shows that performance of 401 kindergarten children on a picture-naming task is the second best predictor of developing difficulties in reading. Similarly, Katz (1986) compares the performance of good, average and poor readers on a picture-naming task and shows that better performance on a picture-naming task correlates with better reading abilities. Studies on the relationship between naming and reading abilities continue to emerge in the literature today. Wood, Hill, Meyer, and Flowers, (2005) reveal that picture naming ability accounts for 76% of the variance in school-aged children's reading scores and, more recently, Araújo, Reis, Petersson, Faísca, (2015) show a strong correlation between naming ability and reading performance in a large meta-analysis. Picture naming was also found to be a predictor of reading comprehension in children (Catts, Herrera, Nielsen, & Bridges, 2015).

In the clinical setting, picture-naming tests are almost routinely administered as part of neuropsychological or psycho-educational evaluations. A body of research

shows that screening for naming abilities is important to rule out dyslexia (Snowling, Wagtendonk & Stafford, 1988), autism spectrum disorder (ASD) (Luyster, Kadlec, Carter, & Tager-Flusberg, 2008), attention deficit hyperactivity disorder (ADHD) (Giddan & Milling, 1999), social and emotional problems (Gertner, Rice & Hadley, 1994; Tervo, 2007) and cognitive impairments (Oliver, Dale & Plomin, 2004). It also predicts reading and spelling abilities in children 24 months after traumatic brain injuries (TBI) (Catroppaa & Anderson, 2004).

In adult populations, a low performance on naming tasks could be associated with neurocognitive or neurodegenerative disorders. Anomia, which is defined as a deficit in naming due to damage to the language areas in the brain, is a symptom characteristic of a neurocognitive disorder (Martin, Fink, Renvall, Laine, 2006). In fact, research shows that anomia is one of the first symptoms of Alzheimer's disease (Huff, Corkin, & Growdon, 1986; Zec, 1993), aphasia (impairment of language; Helm-Estabrooks & Albert, 2004), fronto-temporal dementia (Weder, Aziz, Wilkins, & Tampi, 2007), post-epilepsy surgery (Ives-Deliperi & Butler, 2012), and fragile X syndrome (Spinelli, De Oliveira Rocha, Giacheti, & Richieri-Costa, 1995).

Given the large existing evidence on the relationship between naming ability and other cognitive functions, it is safe to say that evaluating naming performance in the clinical and educational setting is indispensable to understanding individual abilities, documenting language development including fund of vocabulary, and informing appropriate interventions.

Statement of the Problem

Historically, there has been a challenge in evaluating vocabulary knowledge in culturally diverse populations. Given that vocabulary is a reflection of the knowledge

and language use of communities, it is by nature culturally specific (Champion, Hyter, McCabe & Bland-Stewart, 2003). Unfortunately, currently used standardized tests in Lebanon, including picture-naming tests, are developed and normed in Western and European countries (Simhairi, 2010). These include, but are not limited to, the Boston Naming Test - Second Edition (BNT-II) (Kaplan, Goodglass, & Weintraub, 2001), Expressive One-Word Picture Vocabulary Tests, Fourth Edition (EOWPVT-4) (Martin & Brownell, 2011), the Peabody Picture Vocabulary Test, Fourth Edition (PPVT™-4) (Dunn & Dunn, 2007), the Expressive Vocabulary Test, Second Edition (EVT-2) (Williams, 1997) and the Bridge of Vocabulary: Evidence - Based Activities for Academic Success (Montgomery, 2007). Consequently, two issues arise from administering western tests on non-western populations. The first issue pertains to the *cultural relevance* of the stimuli, as items selected on those tests are pictures of objects that are specific to the source culture. For example, the Boston Naming Test, which is the eighth most commonly used test among neuropsychologists around the world (Camara, Nathan & Puente, 2000), includes some picture items that are not familiar to children living in the Middle-East region such as a pretzel, an igloo and an otter. The second issue that arises from administering western tests on non-western populations pertains to the *normative sample* of the test during standardization, as obtained scores of the examinee are compared to a normative sample of people from a different country and culture. For instance, a study shows that the average performance of young adult Spanish/English bilinguals on the BNT is significantly lower than published norms of monolingual English speakers (Kohnert, Hernandez, & Bates, 1998). Those findings have been replicated with different English speaking bilingual groups showing consistently that the average score of minority groups on picture-naming tests is lower

than the average score of the normative sample (Gollan, Fennema-Notestine, Montoya, & Jernigan 2007; Roberts, Garcia, & Desrochers, 2002).

The Risk of Test Bias

Despite the sound psychometric properties of tests when they're administered to their intended population, they evidently introduce biases when administered in other cultures. Bias refers to errors in the validity of the test's results that overestimate or underestimate the values being measured (He & van de Vijver, 2012). It can stem from the test itself (e.g. test items), the aspects of administration (e.g. ethnicity of the administrator) or the normative sample to which the results are compared. Test bias can lead to unreliable decisions particularly when used in "high-stake" conditions such as informing diagnosis or placement. In recent years, researchers proposed several solutions to resolve the issue of test biases in cross-cultural assessments and to allow more individuals to have access to a larger number of assessment tools (Van de Vijver, & Tanzer, 2004). A test can undergo *direct translation* (also called *adoption* or *application*), which is restricted to moving the test from one language to another to preserve its linguistic meaning, or it can undergo *adaptation*, which should provide evidence of semantic equivalence across both cultures and adequate psychometric properties (International Test Commission Guidelines, 2010). Adaptation adds to the process of direct translation in that it could involve changing some items and the creation of new items. Although test adaptation could resolve some of the issues related to test bias, some test items are not transferrable across cultures and have no equivalent in another language (Peña, 2007). In such cases, the test may undergo extensive adaptation to the point where a new test is practically developed. This method, referred to as *test assembly*, is the third method suggested by Van de Vijver, and Tanzer, (2004)

and it is the method we choose to implement in the development of the Lebanese Picture Naming Test.

Emic and Etic Approaches to Test Development

When choosing a method of test translation to implement, researchers should understand the implications it has on future test use and score comparison. A recurrent theme in cross-cultural assessments is the emic and etic perspective to test development. Schaffer and Riordan (2003) describe the emic approach to test development as a method of developing items that measure constructs examined from within the source culture itself and that may not be generalizable to other cultures. On the other hand, an etic approach will examine shared constructs that exist across cultures in the same way to allow for comparative analysis between individuals of different cultures. In this study, we aim to develop a test that is specific to Lebanese children through an emic approach to test development in order to obtain a valid and reliable measure of naming ability in Lebanese children. The test will contain items that are specific to the target culture and may not achieve generalizability across cultures. Given the bilingual and trilingual nature of the Lebanese dialect, one method of test assembly could potentially address the biases and overcome limitations of other methods of test adaptations: the dual-focus approach.

Practical Framework for Test Construction: the Dual-Focus Approach

We will adopt a model for test development that is based on the dual-focus approach developed by Erkut, Alarcón, Coll, Tropp, and García, (1999). The dual-focus approach implements a concept-driven approach that involves a team of professionals who are both indigenous bilingual (or multilingual) researchers and experts on the content of the test (in our case language development in children) in order to minimize

linguistic bias and develop a test that is culturally and linguistically suitable for the target population (Erkut et al., 1999). The method assumes an emic approach to test development as items are chosen based on their relevance to the target culture. Given the multilingual nature of Lebanese children, responses to picture items may be in any of the locally spoken languages (Arabic, French or English). Therefore, test items will be developed in several languages simultaneously in order to render the test suitable to the target population. The original dual-focus approach model comprises five steps that are delineated in Figure 1. In this study, the development of the picture-naming test is based on the original dual-focus approach and may include modifications to some of the steps.

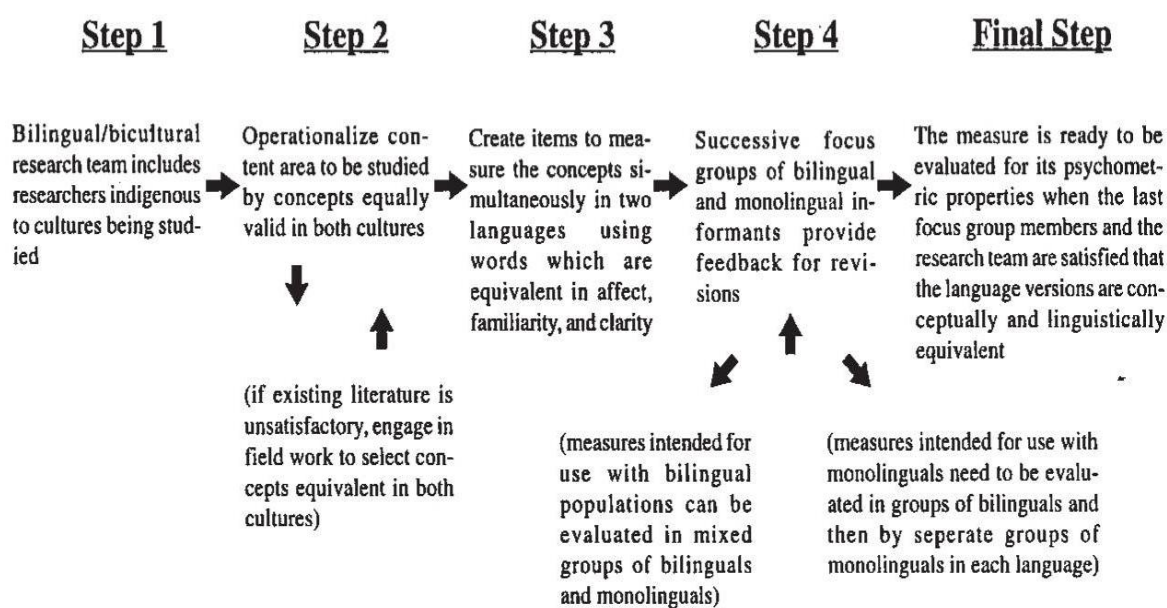


Figure 1. Steps in Creating a Bilingual Measure Using the Dual-Focus Approach (Erkut et al., 1999).

The Need for a Lebanese Psycholinguistic Database

In order to construct a robust picture naming test with good item functioning, the compiled picture words must have known and quantifiable characteristics to allow a sound interpretation of the individual's performance while controlling for properties

that are specific to the words. This is achieved through the development of a *psycholinguistic database* that entails the standardization of words on key psycholinguistic variables such as word frequency, cultural familiarity, age of acquisition and name agreement. A psycholinguistic database is usually developed through collecting data from a sample of individuals who will rate the pictures accordingly or through a screening of text corpuses to collect data on word frequencies and other variables. Unfortunately, a psycholinguistic database that is specific to the Lebanese culture currently does not exist, which poses yet another limitation to the development of a Lebanese naming test (more information on psycholinguistic databases is provided in the study's literature review).

During the recent years, regional efforts led to the development of psycholinguistic databases of Arabic nouns in the Middle East and North Africa (MENA) region. Databases were developed in varieties of the Arabic language such as Levantine Arabic (Lebanon, Jordan, Syria and Palestine) (Khwaileh, Body & Herbert, 2014), Modern Standard Arabic (Boudelaa & Marslen-Wilson, 2010), Tunisian Arabic (Boukadi, Cirina & Wilson, 2015), Saudi Arabic (Alyahya & Druks 2015) and Persian (Bakhitar, Nilipour & Weekes, 2013; Ghasisin, Yadegari, Rahgozar, Nazari & Rastegarianzade, 2015). However, given that Arabic variations are different in terms of word lexicon and pronunciation, the available databases may not be suitable for the Lebanese Arabic speakers who use a considerably different dialect composed of three languages: Arabic, English and French (Encyclopedia Britannica, 2011).

The Current Study

Purpose of the Current Study

The purpose of the study is to develop a draft picture-naming test that is culturally and linguistically suitable for Lebanese children using words with known psycholinguistic characteristics. The study will achieve its purpose through the three objectives listed below:

Objective 1: Generate a Psycholinguistic Database

- 1- Develop the first Lebanese database of psycholinguistic variables with data on word frequency, name agreement, age of acquisition and cultural familiarity, based on experts' ratings of a set of picture words.

Objective 2: Test Development and Test Piloting

- 1- Using data from the database of psycholinguistic variables, select pictures to develop a preliminary draft picture-naming test suitable for Lebanese school-aged children using the dual-focus approach for test development to include items in Arabic, French and English.
- 2- Examine the appropriateness of the developed stimuli on a representative sample of school-aged Lebanese participants between the ages of 3 and 9 years through a pilot study.

Objective 3: Examine the Test's Psychometric Properties and Suggest Revisions

- 1- Conduct analysis at the sample level:
 - a. Run a descriptive analysis to compare group performance based on age, gender and type of schooling (private schools and public schools).
- 2- Conduct analysis at the item level:

- a. Conduct a test item analysis based on Classical Test Theory (CTT) to examine the item parameters: Item Difficulty Index and Item Discrimination Index.
 - b. Evaluate the test's internal reliability.
 - c. Make decisions to retain or revise test items based on item parameter and experts' qualitative comments.
- 3- Compile the remaining items in a second draft picture-naming test which will be referred to as the Draft Lebanese Picture Naming Test (LPNT).
 - 4- Suggest further modifications and revisions to the draft test based on results derived from the pilot phase and test analysis.

Outline of the Methods and Procedure

The section below will briefly describe the three phases of the current study.

Phase 1: Development of a Psycholinguistic Database. A group of indigenous Lebanese experts in the fields of language development, childhood education, special education and neuropsychology will rate a set of pictures on the variables of cultural familiarity, word frequency and age of acquisition between 3 and 9 years old and provide a name for each picture in Lebanese colloquial (spoken) Arabic, and either French, or English or both simultaneously. The experts will also be asked to add qualitative comments to pictures they consider to be poorly illustrated.

Phase 2: Test Development and Test Piloting. Based on the ratings, the committee will discard pictures that are (1) low on cultural familiarity (2) low on name agreement and (3) have an age of acquisition older than 9. The remaining pictures will be compiled and ordered in a logical and developmental sequence based on age of acquisition to obtain the first draft of the picture-naming test. The pictures will be assigned the modal

names (provided name with the highest frequency) provided by the experts in Lebanese Colloquial Arabic, French and English. The researcher will pilot the test on a representative sample of children to whom the measure is intended. Total scores will be tabulated for each participant.

Phase 3: Test Analysis and Suggested Revisions. Using the data collected during the piloting phase, a descriptive and comparative analysis of group performance will be conducted based on age, gender and type of schooling. The test's internal reliability will be examined. Classical Test Theory analysis will involve calculating item parameters. Items with poor item discrimination across three age groups (3-5, 6-7 and 8-9) will be discarded from the picture set. Items with good item parameters but that were considered by the experts to have poor picture illustrations will be examined by two new blind experts to decide on item retention or revision. Modifications and revisions to the test items will be made based on the results of the piloting phase.

Participants

Typical sample size in pilot studies of picture-naming test development ranges from 30 to 70 participants. Fiez and Tranel (1997) collected data from 40 undergraduate psychology subjects during the piloting phase, whereas Panjwani (2012) and Casas et al. (2008) gathered data from a pilot sample of 32 and 67 participants respectively. Our pilot study aims to recruit over 70 typically developing Lebanese children between the ages of 3 years and 9 years 11 months enrolled in regular private and public schools across Beirut and selected according to stratified random sampling methods. Exclusion criteria include pre-existing diagnoses of neurodevelopmental disorders, including language disorder, intellectual disabilities, learning disability and physical disability affecting sensory modalities.

Test and Item Analysis

Analysis at the Sample Level

The pilot sample remains too small for drawing conclusions regarding group differences in test performance. Nonetheless, group comparisons will be made to explore how children from different groups are performing and whether results are consistent with literature on the relationship between the variables age, gender and socio-economic background and naming ability. Previous studies show that older age and higher socio-economic status are both contributing factors to better performance on vocabulary tests (Hoff & Tian, 2005; Ilkman, 2015; Spinelli et al., 2005). Studies on gender remain inconclusive, with some studies showing better performance in males and others in females (Grabowski, Damasio, Eichhorn, & Tranel, 2003; Pineda et al., 2000). Therefore, we expect that, to a certain extent, results of the comparative studies between groups will resemble findings reported in the literature.

Analysis at the Item Level

In order to evaluate the homogeneity of the test items, we will compute measures of internal reliability across age groups and across the total pilot sample. Additionally, given that this is the first step in the development of the Lebanese Picture-Naming Test, it is important to evaluate how individual test items are functioning in order to inform future modifications to the test draft. Item analysis will involve calculation of *item difficulty indices* (the proportion of examinees answering the item correctly) and *item discrimination indices* (the extent to which success on an item corresponds to success on the total score) across three age groups (3-5, 6-7, 8-9). Decisions will be made to discard and retain items based on their item parameters. We will first discard items with poor item discrimination across all three age groups. These

items fail to discriminate between high performing and low performing students.

However, if an item with poor item parameters is assigned an age of acquisition of 3 years old or below, then we will decide to retain the item: the reason being that these items help in expanding the floor of test items and establish a low basal in future test revisions. Remaining items with good item parameters but that were considered by an expert to have poor quality of picture illustration will be evaluated by two new experts that are blind to the item parameters. The experts will either decide to retain or revise and make changes to the picture. At the end of the selection process, the remaining items will be compiled to form the draft LPNT.

Significance and Need of a Lebanese Picture-Naming Test in Lebanon

In the school and clinical setting, naming tests are routinely administered by teachers, special educators, speech and language pathologists, neuropsychologists, clinical psychologists and pediatricians who are required to assess language development in children, implement interventions or refer to recommended services and track progress. Despite professionals being aware of potential biases of imported tests, they continue to administer them regularly for clinical or research purposes, mainly due to the scarcity of locally adapted and developed assessment tools (Simhairi, 2010). The significance of this study lies in its effort to provide local clinicians with a psychometrically reliable tool to assess expressive vocabulary in Lebanese children. To our knowledge, a naming test that is culturally fair to Lebanese individuals is not yet developed nor adapted. This test could become the first standardized measure of expressive vocabulary that is fair to Lebanese children and that can provide clinicians with rapid, robust and reliable results. It is also the first attempt to develop a Lebanese psycholinguistic database of picture words to be used in future research.

Assumptions and Foreseen Limitations

The study makes the assumption that naming ability in children is measured and operationalized through a task of picture naming. It also assumes that the piloting sample is somewhat representative of Lebanese students aged 3 to 9 years in Beirut's third district area and that test administration during the pilot phase will follow standardized procedure and will be minimally affected by extraneous variables such as the administrator's characteristics or method of instructions delivery. Finally, it assumes that the committee of experts is composed of reliable professionals and that the picture ratings assigned are based on a professional and educated judgment. The study also foresees several limitations. The picture set used may not represent the best choice of items for the test and the researcher may later consider having the pictures illustrated from scratch in future drafts. The size of the pilot sample may not be enough to draw conclusions from comparative studies or item functioning analysis and should be enlarged in future piloting studies. A convergent validity study will not be carried out since there is no available picture-naming test for Lebanese children and any other tool may introduce bias. Upon termination of the study, additional limitations will be revealed and discussed.

Summarizing the Current Study

Naming is a measure of expressive vocabulary that plays an important role in predicting cognitive functions. Currently used tests to assess naming abilities in Lebanese children are imported from Western countries and pose a threat to the validity of test results. It is now established that translated and adapted tests do not eliminate cultural biases and may introduce the possibility of having the construct assessed altered when changing items. The ideal solution would be to create and develop from scratch

construct-equivalent tests relevant to the Lebanese culture. Because a psycholinguist database of picture words in Lebanese Arabic is currently nonexistent, we will also develop the first database through experts' rating of psycholinguistic variables in order to choose the pictures based on standardized variables. Our test construction method will adopt the dual-focus approach to develop items in Arabic, French and English simultaneously. The first draft will be piloted on a representative sample of Lebanese children between the ages of 3 and 9 years. Results from test analysis will be used to retain and Review test items to finally compile a second-draft picture-naming test.

Chapter 2: Literature Review

The following chapter will present the large literature on naming ability and naming tests in the context of cross-cultural psychology. The chapter is divided into four parts, which are all of equal relevance to our study. The first part provides an overview on language development in children, lexical retrieval models and defines naming as a measure of expressive vocabulary. The second part introduces the unique Lebanese dialect with an emphasis on its multilingual nature. The third part presents the different types of testing biases, test translation and adaptation methods including the dual-focus approach, which we are adopting in the construction of the Lebanese picture naming test. We complete the chapter with the fourth part on recent research on naming in the Arab World, the development of Arabic and multicultural psycholinguistic databases and a description of the MultiPic Database used in this study. Concerning terminology, we will often use the word *tests* to refer to instruments, inventories, questionnaires, schedules, assessment tools or scales.

Part 1: Language Development, Naming and Lexical Retrieval Models

First Words

Language development is the process by which children come to understand the world and begin to communicate with others. Not surprisingly, communication in infants begins way before they have acquired any words. At 12 months, infants show communicative motives by pointing and using word utterances (word fragments) to inform others of a need, and at 18 months, they begin to combine two words to make a sentence (Tomasello et al., 2007). Between the ages of two and three years, children begin to form early abstract sentences such as questions and imperatives (e.g. *Where Daddy going?* or *Push here*). They go on to develop increasingly complex sentences

with expected errors of grammar and syntax during the early stages on speech. As language abilities develop, children's vocabulary begins to expand. By 24 months, a child's vocabulary reservoir is close to 200-300 words that are usually names of common everyday objects. Vocabulary increases to reach 900-1000 words by 3 years, 1500 words by 4 years, 1500-2200 words by 5 years, 2600 words by 6 years and more than 50 000 words by adulthood (Owens, 1984), a fourth of which consists of names of objects (Levelt, Roelofs & Meyer 1999).

Vocabulary is also a factor that emerges in larger models of cognitive functions such as the Cattell-Horn-Carroll (CHC) theory- an empirically supported psychological theory on the structure of human cognitive abilities. The theory delineates, through confirmatory factor analysis, the relationships between variables, how they influence each other (Keith & Reynolds, 2010). In the CHC model, Vocabulary is a factor of Crystallized Intelligence, which is defined as the ability to use acquired knowledge and skills that consist primarily of verbal and language-based knowledge accumulated during education and general life experiences (Horn & Blankson, 2005). The term crystallized suggests that this type of knowledge has become frozen and consolidated (Cattell, 1987). This implies that vocabulary, although defined as a separate measurable construct, also spills into other linguistic and cognitive abilities.

Picture Naming as a Measure of Expressive Vocabulary

There are several ways to measure vocabulary and the method of choice depends on the purpose of the assessment and the type of vocabulary that we wish to assess.

Receptive vocabulary can be measured through a matching task where an individual is asked to match a word with its correct representation (usually a picture), whereas *expressive vocabulary* can be measured through a naming task where an individual is

shown a picture and asked to provide one word that names the picture. The simple format of a naming task makes it easy to administer and practical. Pictures have been used with success in evaluating naming because recognizing a picture of an object does not require learning or development beyond learning the name of the object itself (Glaser, 1992).

An Overview on Lexical Retrieval Models

To facilitate word production in children during a naming task in children, a series of events take place in the brain. These events were a subject of research for a long time and were integrated into several models called lexical retrieval models. Different researchers suggested different models that underlie naming which are in fact somewhat similar. We will briefly describe some of the most prominent lexical retrieval models that emerged from different starting points across the years.

Johnson, Pavio and Clark (1996) described three broad stages that underlie the naming process. In the first stage, an individual examines the object and identifies it as a member of a known category of objects. In the second stage, “name activation” of the object occurs among thousands of words in the individual’s vocabulary reservoir and in the third and final stage, articulatory commands for the specific word result in the generation of a response. These stages occur rather sequentially and rapidly in fluent speech. Later studies resonated with similar models of naming. Dell, Schwartz, Martin, Saffran, and Gagnon (1997) described the process as beginning with the conversion of the visual stimulus into a conceptual representation in the brain, followed by the retrieval of the appropriate name and the corresponding mental picture and ending with an articulation of the name. Shortly after, a model of lexical retrieval emerged that remains until this day the standard reference model. In 1999, Levelt, Roelofs and Meyer

developed a 5-stage model for lexical production. It begins with (1) conceptual activation, which refers to the activation of a category of words as a response to a visual stimulus, (2) lexical selection, which consists of retrieving a word from a mental lexicon that contains tens of thousands of items, (3) morphophonological encoding and syllabification of the word, which makes the move from the conceptual or abstract word to be retrieved to the actual morphology, and phonology of the word (letters and sounds). In the fourth stage, the individual undergoes phonetic encoding which is a preparatory phase for the articulation task that will produce the word, (5) and finally, articulation of the word where the coordinated movements of the muscles of the lungs, larynx, and vocal tracts are executed. Evidently, naming is a complex task that taps all the stages of lexical production, lexical access, phonological coding, and articulation muscle coordination stored in the individual's memory. Deficits in lexical retrieval have been attributed to several causes some of which are related to poor storage (Dollaghan, 1987) or phonological processing (Constable, Stackhouse & Wells, 1997).

Although language is considered a universal mode of communication, there exist about 6000 to 7000 languages in use today (Northrup, 2005) many of which also include a variety of dialects. Arabic is the language of interest to this study as it is the official national language in Lebanon. However, over the years, not only did the Lebanese population develop an Arabic dialect that is unique to the population but also integrated two other languages for daily communication use. The languages and dialects molded into each other to create what could possibly be called at this point: The Lebanese dialect. The next part of this chapter will delve into the details pertaining to the Lebanese dialect and its unique features.

Part 2: Lebanon, the Case of a “Unique Multilingual and Multicultural Make Up”

Introducing the Lebanese Dialect

Lebanon is considered to have a “unique multilingual and multicultural make up” (Bacha & Bahous, 2011). Lebanon’s official national language is Arabic with the majority of Lebanese people using Spoken Lebanese Arabic Vernacular (derej or دارج), part of the Levantine Arabic, as the primary mode of communication in daily conversations. Lebanese are also exposed to Modern Standard Arabic (MSA or فصحي) as the official language in magazines, news broadcasting and newspapers. Lebanese Arabic is a classic example of diglossic language where two varieties of the language are used under different conditions in the same community (Ferguson, 1959). Despite some similarities in several linguistic features of both dialects, Spoken Lebanese Arabic Vernacular and MSA have distinct syntactic, morphological, phonological and lexical characteristics (Holes, 2005; Versteegh, 1997).

On top of being exposed to two different varieties of the same language, and sometimes a variety that lies somewhere in between, most of the Lebanese people are also bilingual or trilingual. As a result of a cultural exchange on Lebanese territories in the 20th century, French and English languages became second languages of instruction in Lebanese schools. Lebanon is reported to have the highest literacy rate amongst its neighboring countries in the Middle East with a literacy level reaching 91% of adults and 99% of youth (UNESCO, 2009). Statistics from the Lebanese Ministry of Education in the year 2018 show that 51.4% of all Lebanese schools offer French as a primary language of instruction and 48.6% offer English as a primary language of instruction (CERD, 2018). Children begin to learn a foreign language, French or English and sometimes both, at the Nursery level in private schools and most public

schools (Shaaban, 1997). French and English language classes are offered on average 8 hours a week at the Elementary level and they are also used as the medium of instruction for mathematics, sciences and social studies at all levels. Furthermore, 20% of the Lebanese population uses a foreign language on a daily basis (Encyclopedia Britannica, 2011) and it is predicted that, with time, the Lebanese youth will continue to use more foreign language in their daily dialect and less Lebanese Arabic (Shawish, 2010). With this being reported, school-aged children in Lebanon are all generally bilingual (Arabic-English or Arabic-French) and sometimes trilingual.

Lebanon's "Two-Faced" Educational System

Lebanese schools are referred to as either "public school" or "private school" depending on the sector to which they belong. Schools in the public sector are financed by The Ministry of Education and Higher Education (MEHE), whereas schools in the private sector are generally financed by students' fees. Unfortunately, spending on education in the public sector consistently falls short of spending on education in the private sector. According to Lebanon's National Accounts, an analysis of Lebanon's education expenditure shows that during the year 2011, the sum of expenditure on education in the public sector was approximately USD 641 million, which was equal to 1.6% of Lebanon's GDP during that year, whereas it reached almost the triple in the private sector, USD 1,783 million. Compared to other Arab countries, spending on public education in Lebanon is significantly low with countries like Tunisia spending 6.2% of their GDP on public education and KSA 5.6% (Soueid et al., 2014). Additionally, reports show that students enrolled in private schools have a higher success rate in intermediary exams (83.1% compared to 64% in public schools), more qualified teachers and personnel and are less likely to repeat a grade than students

enrolled in public schools (PNUD Report, 2009). Reasons for the differences in the quality of education provided in private schools and public schools are also the result of decades of political and sectarian conflicts that are over and beyond the scope of this section despite their implications being drastic on Lebanese students and education till this day. Although there are no reliable reported numbers showing the difference in socio-economic status of children enrolled in private schools compared to those enrolled in public schools, it is believed that Lebanon's educational system became categorically divided into private sector and public sector, where families belonging to middle to upper income groups are enrolled in the former and families from lower social-economic backgrounds are enrolled in the latter (Frayha, 2009).

Implications of Bilingualism on Language Testing

As mentioned, Lebanese children are dominantly bilingual and sometimes trilingual. Until recently, bilingualism in children was considered a unique phenomenon rather than a typical one (Crystal, 2004). However, bilingualism is becoming more common than it used to and research in this area has been increasing with the number of articles on bilingualism almost tripling in the literature between 1997 and 2005 (Bialystok, 2007). Accordingly, theorists now suggest that Chomsky's model on the Language Acquisition Device (LAD; Chomsky, 1965) is in fact a Multilingual Acquisition Device (Crystal, 2004), which can be defined as an innate ability in children to acquire several languages. Pearson (2008) identifies four types of bilinguals: (1) *active bilinguals* are individuals who can produce and understand novel sentences in both the first and second language, (2) *elective bilinguals* are individuals who decide to learn a second language but still use their first language for communication, (3) *immigrant bilinguals* are individuals who move to a new environment or culture and

must learn a second language for their daily livelihood and (4) *passive bilinguals* are those who can only understand or read a second language but cannot produce it. It is possible to say that Lebanese children are considered active bilinguals given that they spend about 7-8 hours at school communicating in the class setting in a second language. Another distinction between bilinguals relates to the use of the languages. Cummins (1979) describes independent versus. interdependent development of the first and second languages. An independent bilingual possesses two independent language system that develop in parallel with minimal overlap also described to have two sets of mental furniture (Wierzbicka, 2005), whereas an interdependent bilingual, similar to a Lebanese child, has one set of mental furniture with two different labels on each piece, each in a different language. Grosjean (2001) argues that bilinguals do not function as “two monolinguals in one person” but as one monolingual who can continuously switch between being a monolingual speaker of each language. Others, do not treat their both languages as separate, and naturally use both languages together in a “bilingual mode” (Gupta, 2006). This leads to unique phenomena called *code switching* and *code mixing* with the former occurring when bilinguals switch languages between sentences and the latter within sentences. In many countries in the world such as India, Singapore, and evidently Lebanon, code mixing is a typical style of communication with speakers switching back and forth between two languages sometimes due to filling in words that are difficult to recall in one of the languages. Another interesting observation is when a word or phrase from one language becomes embedded within a sentence in another language and takes the order and morphosyntax of the other language while molding into expressions and phrases where both languages are used within the same sentence. This observation applies to the Lebanese dialect, where individuals often mix two or

three languages in one sentence during conversations. Because the Lebanese dialect is particularly unique, any assessment of language in Lebanese children should in fact reflect the language in use. Unfortunately, in school settings or clinical settings, professionals who assess language skills in children typically use standardized tests normed against English or French monolinguals that exclude children who are bilinguals from the normative samples because they may skew the results (Gathercole, 2013).

In a text on Solution for Assessing Bilinguals, Gathercole (2013) states that speech and language therapists face issues when working with bilingual children due to the diversity within the bilinguals, the lack of test materials and norms, and the lack of knowledge regarding the characteristics and language development of bilinguals. She continues to describe the few available tests that are specifically developed for bilingual children such as the *Prawf Geirfa* (Welsh-English vocabulary test, Gathercole et al. 2008), Sandwell bilingual screening assessment scales for Punjabi and English (Duncan et al., 1988) and the test for auditory comprehension of language English/Spanish (Carrow, 1973) and highlights on the need for more tests suitable for bilingual children.

Given that tests of language used in Lebanon are not testing the Lebanese dialect, there is a genuine need to develop tests that are linguistically suitable for Lebanese children. Any assessment of language or communication abilities should take into account the complex relationship between spoken dialects of Arabic, MSA and other learnt second languages. This implies that developing a naming test that permits answers either exclusively in Arabic or English or French may not provide us with an accurate estimation of the child's naming ability. In the current study we adopt an approach for test construction that allows the development of test items in multiple

languages simultaneously in order to reduce the amount of testing bias and obtain a fair and sound measure of vocabulary in Lebanese children: the dual-focus approach for test development.

Part 3: Testing Biases and Implications on Assessments of Vocabulary

The third part is of particular relevance to our study as it sheds light on the threatening biases of currently used tests in Lebanese schools and clinical setting. Here we present the different types of testing biases suggested by Van de Vijver (1997), followed by methods of test translation, adaptation and assembly to reduce testing bias and completing the section with an overview on the dual focus approach method for test development which we adopt in developing our test. The review covers broad literature on cross-cultural assessments and is *not* limited to language tests since issues in cross-cultural assessment of broader cognitive abilities also apply to language assessment.

There exists a long-standing history of challenges in measuring vocabulary in culturally diverse populations and vocabulary tests have been a subject under study in the field of cross-cultural assessments for a long time. Before delving into the types of testing biases, we present a brief overview on the field of cross-cultural assessments to date.

Background on Cross-Cultural Assessment and the Emergence of Test Bias

An increasing amount of research is showing that culture and psychology are no longer considered two distinct fields of study. It is now well established that culture has a pervasive influence on individual human abilities and it is an important variable in all aspects of human psychology (Bond, Van de Vijver, & Matsumoto, 2011). Cross-cultural psychology is defined as the study of the interplay between culture and psychological variables (Georgas, 2003). It is an inter-disciplinary field concerned with

the commonalities and variations of human psychological processes across different cultures and their implications on research, practice and assessment (Berry, 2002). Particular interest in cross-cultural assessment emerged due to increasing societal concerns regarding the administration and interpretation of different types of assessments in cultures they are not designed for (Puente et al., 2013). An early research estimated that more than 5 million students are tested every year by standardized achievement tests inappropriately due to differences in cultural backgrounds (Torres, 1991). In fact, research also showed that children from cultural minorities understand test items differently and score lower than children of the mainstream culture on standardized assessments (Gopaul-McNicol & Brice-Baker, 1998). Differences in patterns of performance on cognitive ability tests due to cultural backgrounds were reported in several domains of functioning including intelligence tests (Kaufman, 2009) and picture naming tests (Serpell & Deregowski, 1980). To illustrate, an early study by Lieblich (1983) showed that children of Asian-African origins score 13 to 15 points lower than children of European-American origin on intelligence tests. Similarly, Reynold and Ramsay (2003) reported that African-American minorities score 1.0 standard deviation lower than Whites in the United States on intelligence tests. Such provocative results lead researchers and practitioners to raise questions about the validity of scores and potentially attribute low scores of cultural minorities to poor testing practices. The misuse of assessment tools is an issue that holds long-term and sometimes permanent consequences for children, clients, and parents. For young students, misinterpretation of test scores may lead to misdiagnosis and major groundless decisions regarding placement in the educational setting. For older individuals, the consequences can range from college rejection, to denial of employment and

sometimes, legal actions (Reynolds, 2000).

Concerns over cultural bias in testing were also addressed in official documents such as the American Psychological Association Ethics Code. The APA Ethics Code states the psychologists should interpret assessment data carefully while “*keeping in mind the cultural and linguistic characteristics of the individual being assessed*” (APA Ethics Code, Standard 9.06, p. 13). Moreover, in 2016, the International Test Commission released the latest version of Guidelines for Translating and Adapting Tests (Second Edition) in an attempt to establish score equivalence across different cultures. The document lists 22 guidelines under test administration, test development, scores interpretation and context in order to minimize testing bias to the largest extent possible (International Test Commission, 2016).

Bias in Cross-Cultural Assessment: Definition and Types

Bias refers to errors in the validity of test results that overestimate or underestimate the values being measured (He & van de Vijver, 2012). A biased test score may therefore not adequately reflect constructs, traits or abilities across different cultures. Based on the bulk of research showing evidence of bias in cross-cultural testing, Van de Vijver (1997) developed the first taxonomy depicting three different sources of testing bias across cultures: construct bias, method bias and item bias. The following section will describe the types and subtypes of testing bias and provide relevant examples.

Construct Bias. Construct bias occurs when the same construct is not defined and perceived similarly across cultures. For example, non-Western societies’ conception of intelligence differs from Western societies’ conception of intelligence in

that it involves social relationships and interpersonal skills in addition to the scholastic domains (Serpell, 1993; Super, 1983).

Methods Bias. Methods bias branches into three subtypes of bias: sample bias, administration bias and instrument (test) bias.

Sample Bias. Sample bias occurs when selected samples cannot be compared due to underlying differences in abilities or education. For example, students who did not receive any form of education cannot be compared to a sample of students who received formal education. Similarly, Lebanese children cannot be compared to children from a Western society.

Administration Bias. Administration bias occurs when testing instructions, expertise of administrators, language and communication are different across cultures. Several studies found differences in test scores when the test administrator behaved in different ways. For example, studies found that individuals receive better scores when the administrator of the test was of a similar cultural background (Little & Ramirez, 1976) or portrayed positive nonverbal behavior during testing (Saigh, 1980). Administration bias also involves differences in the preferred testing conditions and response procedures across cultures such as paper-and-pencil tests versus online surveys (Dwight & Feigelson, 2000). For example, Zambian children perform better than British children on tasks that require pattern reproduction if they are allowed to use iron wires to reproduce the model instead of a paper and a pencil (Serpell, 1979).

Instrument Bias. Method bias also includes instrument bias, which refers to differences in the familiarity of the stimulus material used for testing. A commonly cited example is that of Piswanger (1975) where Arabic speaking students were compared to Austrian students taking the figural inductive reasoning test. Results of the

study showed that Arabic speaking students performed poorly on tasks demanding the application of rules in a left-to-right horizontal direction. The results could be attributed to the differences in writing direction between both cultures (Piswanger, 1975).

Item Bias. The last source of bias according to Van de Vijver is item bias (1997). Item bias refers to a differential in the psychological meaning of an item across different cultures. In other words, an item is biased if it is answered differently by two individuals from different cultures with similar traits. This is often due to poor item translation, differences in the familiarity of the item across cultures or ambiguous wording of an item (Van de Vijver, 2004). For example, some English language expressions and idioms cannot be literally translated to Arabic or any other language while holding the same meaning such as “I am feeling blue”. Item bias is of particular relevance to picture naming tests where pictures of objects that are common in one culture may be unfamiliar to individuals from another culture. For example, on the Boston Naming Test, a score of zero on the items showing an igloo, a pretzel or an otter to a Lebanese does not necessarily reflect a weakness in naming but possibly unfamiliarity with the picture presented due to environmental, societal and cultural differences. Test users need to pay careful attention to the sources of bias to avoid contaminating test results and assess children’s true abilities. In the following section, we will discuss suggested ways to reduce test biases and methods that researchers adopt to move a test from one culture to another.

Current Practices in Translating and Adapting Multicultural Assessment

Translating tests is a process that goes beyond mere rewriting of the items in another language. Its purpose is to facilitate comparative cultural studies, allow individuals to be tested in their own language and also reduce spent resources to

develop completely new tests (Reynolds & Suzuki, 2003). The International Test Commission (ITC) released guidelines for test translation and adaptations with the latest version published in 2016 to improve the quality of test adaptation practices across cultures. The guidelines cover six broad topics: Pre-Condition (3), Test Development (5), Confirmation (Empirical Analyses) (4), Administration (2), Score Scales and Interpretation (2), and Documentation (2). Since their first edition, the guidelines remain a central reference for test adaptation. This study will follow the guidelines suggested to ensure sound quality test construction using a best-practice approach.

In *The Handbook of Multicultural Assessment*, Padilla and Medina (2001) list several reasons for which translation of tests is complex to many researchers. They state that test directions are frequently too “psychotechnical” (p. 20) to allow for direct translation and that the tested psychological constructs are not always universal. They further add that the examinee’s testing behavior also varies across cultures. Up till 1980, common standards for translating tests were not yet reported (Brislin, 1980) and researchers were calling for a comprehensive multistep translation process in order to develop normative interpretation of assessment (Giensinger, 1994). In *Methods and Data Analysis for Cross-cultural Research*, Van de Vijver and Leung (1997) proposed for the first time three ways to translate an existing test to be used in multilingual settings. Before describing the method adopted to develop the draft Lebanese Picture Naming Test, we will review other existing methods of translation namely adoption of a test, adaptation of a test and the assembly of a test.

Adoption: Direct Translation. The first option for test translation, which is the most common, is called adoption. Adoption involves the direct translation of items in a test. There are two reported ways to translate a test. The first being *forward-backward*

translation which involves a forward translation of a test from the source language to the target language, followed by a back-translation of the text by an independent interpreter back to its source language and finally an evaluation of both versions by comparing the original test with the back-translated test. This process has been widely applied in Lebanon and in the Arab region in the translation of different scales, tests and inventories (Berri, & Al-Hroub, 2016; El Hassan & Sader, 2005; El Hassan & Jammal, 2007; Khamis, 2015; Zeinoun et al., 2013). Although adoption of tests is commonly implemented, Hambleton notes that some linguistically accurate translations may not represent similar meanings and may therefore lead to flawed results (Hambleton, 1994). Saeed and Fareh (2006) provide a relevant example to the Arabic language that involves the translation of the discourse marker *fa* into English. They state that such seemingly simple linguistic features result in an improper translation since, according to the results of their study, the same connector *fa* in Arabic can potentially mean “hence”, “therefore”, “because”, “however”, “but”, “then” and “consequently”. Another example relates to tests involving digit span which requires the examinee to repeat a list of numbers verbatim. Puente et al. (2013) point out that the name of some digits in English is shorter than in Arabic such as the number eight (one syllable) and its Arabic translation *thamaniyah* (four syllables), which may make affect attention and retention during this task. A second procedure for translation called *committee approach* may detect such problems (Borsa, Damasio & Bandeira, 2012). The committee approach procedure involves a multidisciplinary team that often includes linguistics, psychologists, and anthropologists, and each contributes to the translation process of tests from a source language to a target language. To implement the committee approach in Arab countries, the committee must include bilingual individuals who are

experts in different domains relevant to the process of test translation and experts in the content area of the test in order to accurately translate the test.

Adaptation: Making Items Culturally Fit. The second option for test translation is called adaptation. According to the International Test Commission Guidelines (2017), the process of adapting tests should provide evidence of semantic equivalence and adequate psychometric properties. It is important to note a distinction between the terms adaptation and adoption. Adaptation goes beyond adoption in that it takes into account elements that concern the cultural fit of an instrument (Hambleton, 2005). Adaptation could involve changing some items and the creation of new items. It may result in changing culturally specific details such as currencies, metric scales and reference to specific places. For example, metric scale changes in translating tests to Arabic would involve changing inches to centimeters and pounds and kilograms. At times, a test undergoes extensive adaptation to the point where a new test is practically developed.

Assembly: Developing New Items. The third option is called assembly. Researchers resort to assembly when the same construct across different cultures hardly overlaps or when several items need to be changed to the point where a new test is practically developed. An indication to use test assembly is when test item content threatens a direct comparison of performance (van de Vijver & Tanzer, 2004). Assembling a test poses a disadvantage to researchers as it prevents comparing the already-existing data to the newly acquired data and it is also more time and resources consuming than test adaptations, therefore the choice of test translation (Hambleton, 1993). However, the study is not so much interested in comparing naming abilities across cultures and developing metric equivalence (full score equivalence) as it is in

working towards developing a tool that can measure naming ability without the interference of confounding variables while maximizing the ecological validity of the instrument. Therefore, we adopt a method of assembly where we practically develop new items in order to obtain a culturally fit test. Given the bilingual and trilingual nature of the Lebanese dialect, one method of test assembly could potentially address the biases and overcome limitations of other methods of test adaptations: the dual-focus approach.

The Dual-Focus Approach

The dual-focus approach for test development was developed by Erkut, Alarcón, García Coll, Tropp, and Vázquez García in 1999 to allow the development of two or more language versions of a test simultaneously. It aims to create bilingual measures where test items are developed in several languages at the same time. In the original study on the dual-focus approach by Erkut et al. in 1999, the authors describe two features that are specific to this approach. First, test construction involves a team of bilingual and bicultural researchers who are indigenous to the test's target cultures. Having a bilingual and bicultural team is also essential to guard against unintended transfer of concepts from one culture to the other during the process of test adaptation (Erkut et al., 1999). Second, it is a concept-driven approach rather than a translation-driven approach to attain equivalence. The involvement of a bilingual and bicultural team in test construction will facilitate the equivalence of concepts and wordings in the data collection protocol. In the original study, Erkut et al. (1999) describe the steps of the dual-focus methodology as follows (refer to Figure 1 in chapter 1):

1. A committee is formed that includes bilingual experts who are familiar with the research topic.

2. The committee members evaluate whether the construct being measured has a conceptual equivalence across the three languages.
3. When a consensus is reached regarding the cross-cultural equivalence of the construct, the third step takes place, which involves the generation of items in different languages simultaneously. If an item does not have a reliable translation in the other language, it is deleted from the set.
4. The researchers receive external feedback from monolingual and bilingual individuals for whom the test is intended.
5. The developed test is piloted.
6. Evaluation of the test's psychometric properties takes place.

In brief, the dual focus approach is a step-wise method of test assembly that is meant to facilitate the development of test items in order to ensure that all items are linguistically and culturally appropriate for the target languages.

Statistical Methods to Detect Bias in Assessment

Statistical methods can be used to estimate the amount of test bias after all items have been administered to a sample. If two samples have similar underlying abilities (vocabulary in our case), then they are expected to score similarly on individual test items unless the item is biased to one group over the other. Test items are considered biased if analysis shows differences on item performance based on group membership alone (and not ability). Ideally, test items should be relatively free of relationships with group membership be it gender or types of education. To illustrate, one can expect that a sample of boys and a sample of girls who have similar naming ability will score differently on pictures of gender-specific toys or activities (*e.g.* soccer, beauty products).

Many statistical methods have been used to detect bias: methods based on the Classical Test Theory (CTT), methods based on Item Response Theory (IRT) and more recently, methods using more computationally complex analysis such as Differential Item Functioning (DIF) which are part of IRT (Lim & Drasgow, 1990). The choice of method depends on the pragmatics of the test-development situation and the stage of development. At an early stage of test development, methods based on CTT are recommended (Martin & Brownell, 2011) to determine the order of item presentation and which items to retain. Additionally, given that this is a pilot study with a small sample size, IRT and DIF may not generate accurate estimations of bias. Therefore, combining item parameters (item difficulty index and item discrimination index) and qualitative comments from the experts will detect items that require careful examination.

Part 4: Psycholinguistic Databases and Naming Tests in the Arab World

Describing Psycholinguistic Variables

In order to use picture items in the development of picture naming tests or for research purposes, researchers should have access to a set of pictures that are standardized across psycholinguistic variables (Alario & Ferrand, 1999; Bonin et al., 2003; Gilhooly & Logie, 1980; Snodgrass & Vanderwart, 1980). Some of the most common standardized psycholinguistic variables are:

1. Age of Acquisition (AoA): the age at which individuals usually learn a given word.
2. Name agreement: the extent to which individuals agree on a single name to refer to a picture.

3. Image agreement: the extent to which mental representations of words match the word that names the picture.
4. Word Frequency: the frequency with which a word occurs in a given language across modes of communication.
5. Familiarity: the extent to which we come in contact with or think about the concept in our everyday life.

Other less used psycholinguistic variables such as visual complexity of the picture, imageability, concreteness of the image, typicality, and naming latency of the words presented. A normative database for picture stimuli and their corresponding nouns is crucial to allow researchers to draw conclusions from a set of data and compare naming performance across datasets.

Psycholinguistic Databases for Arabic Words

In the past 5 years, and in response to an obvious need, psycholinguistic databases emerged in the Arab world. Khwaileh, Body and Herbert (2017) justify the need for such data in Arab populations by highlighting specific characteristics of the Arabic language that do not exist in English language and that that may affect the variables under study. Some of these characteristics are variations in the types of plural words in Arabic (*i.e.* not all plural words in Arabic have similar terminologies), dual plural words (*i.e.* those that refer to pairs), rational and irrational words (*i.e.* words referring to humans and or non-human objects) and gender inflected suffixations that are specific to each gender. Those features, among others, render the results on Arabic naming tests non-comparable to their English equivalents simply because they affect differently an individual's ability to retrieve words. Hence, psycholinguistic databases were developed in Levantine Arabic (Khwaileh, Body & Herbert, 2014), Tunisian

Arabic (Boukadi, Cirina & Wilson, 2015), Persian (Bakhitar, Nilipour & Weekes, 2013; Ghasisin, Yadegari, Rahgozar, Nazari & Rastegarianzade, 2015) and Saudi Arabic (Alyahya & Druks 2015). We briefly present some of these studies and their sample population. Khwaileh, Body and Herbert (2015) identified a set of 186 culturally and linguistically appropriate concept labels and their corresponding photographic representations for Levantine Arabic. Levantine Arabic speaking participants provided norms for visual complexity, imageability, age of acquisition, naming latency and name agreement. The study included 22 participants among whom 12 were Jordanian, 6 were Palestinian, 2 Syrian and 2 Lebanese. In a study by Bakhtiar et al. (2013), normative data pertaining to the ability of picture naming in Persian speaking individuals aged 18 to 29 years was gathered. Ghasisin and Yadegari (2014) extended the norm to include middle-aged and elderly individuals. Normative Data was also developed in Tunisian Arabic for 348 object names (Boukadi, Zouaidi & Wilson, 2015) where they used the line drawings from Cycowicz, Friedman, Rothstein and Snodgrass (1997). Norms were also developed for MSA or “foshā/fos7a” in a study titled ARALEX (Boudelaa & Marslen-Wilson, 2010). ARALEX is a computerized lexical database for MSA based on a contemporary text corpus of 40 million words. However, ARALEX presents psycholinguistic variables of words in MSA and not spoken Lebanese dialect, which differ in terms of use and target population.

Multi-Cultural Psycholinguistic Database

Given that naming tests in different languages will result in different normative data, one would wonder whether it would be possible to develop a test where data is collected from a multi-cultural population so that it is suitable to individuals from different cultures. A few researchers responded to this need through studies and

reviews. Ardila (2007) published a review entitled “Toward the Development of a Cross-linguistic Naming Test” in which the author lists the criteria that a universal naming test should strive to fulfill. Ardila states that the test should include only “universal” words recognized by all languages, which is also known as the “Swadesh word list”. The Swadesh word list was developed in 1952 includes basic universal or core vocabulary that every individual, regardless of time, place and living conditions and culture is expected to have encountered (Swadesh, 1952). It should also include different semantic categories such as living, non-living, and action words, and it should avoid the confounding of perceptual difficulties. Having a cross-cultural naming test presents several advantages such as having the test readily available in all languages, and having the flexibility of changing the photographs and pictures. Another attempt to develop a naming test that could be used across cultures is the Multilingual Naming Test (MINT) by Gollan, Weissberger, Runnqvist, Montoya, and Cera (2012). It consists of a set of 68 black-and-white line drawings selected and presented in order of estimated increasing difficulty. The developers of the tool state that they have carefully designed the test for the assessment of subjects in different languages (English, Spanish, Hebrew, and Mandarin Chinese). However, till this day there are no psychometric data present on whether the tool is valid or reliable and no norms were developed.

The Ideal Picture Set

In order to develop a psycholinguistic database of picture words, we evidently need to start with a set of picture. To date, Joan G. Snodgrass and Mary Vanderwart’s picture set of 260 pictures (Snodgrass & Vanderwart, 1980) is the primary picture set for research on picture naming, with their original study cited over 4750 times (Google Scholar) and normed in tens of languages. However, Snodgrass and Vanderwart’s

picture set presents several limitations, some of which are listed by Dunabeita et al. (2017). For example, the pictures include black-and-white line drawings which have a smaller rate of recognition compared to colored pictures representing the same objects, the picture set is not freely available for use (researchers have to pay to retrieve the pictures), and the picture set consisted of only 260 drawings which requires researchers using the set to add additional pictures from other sources (see the International Picture-Naming Project, IPNP, by Bates et al. 2003 and Szekely et al., 2004). Ideally, and based on comparative studies of performance using different picture sets, a picture set should meet the below criteria (Adlington, Laws, & Gale, 2009; Dunabeita et al., 2017; Rossion & Pourtois, 2004):

- Consistency of drawing style across all pictures
- Colors used in the pictures are the true colors of the object
- Non-ambiguous representations to elude naming failure due to false recognition (high in typicality)
- Animals in pictures are depicted in sideways view
- Objects in pictures are positioned with the functional end towards the bottom (e.g. fork, pencil)
- Printed size of the pictures in decent resolution is at least 10 cm x 10 cm large
- Picture set is larger than 200 pictures and includes pictures from a variety of semantic categories (i.e., objects, furniture, fruits, animals, tools, etc.)

The MultiPic Databank

The MultiPic databank is a recently published picture set of 750 colored line drawings developed digitally by the same illustrator. It is the outcome of a collaborative European project that aimed to create a set of 750 publicly available pictures for the

scientific community as a useful tool for cognitive scientists, researchers and practitioners in the field of language, visual perception, memory and attention in monolingual or multilingual populations (Dunabeita et al., 2017). The database was initially composed of 600 words from the Spanish lexeme database ESPAL based on imageability and concreteness of the pictures (See Duchon, Perea, Sebastian-Galles, Marti & Carreiras, 2013) and the rest of the words were added at a later stage (source of the additional pictures is not specified). The authors took several measures to overcome limitations of other picture sets through expanding the number of pictures included, having one illustrator produce all the pictures to ensure consistency of style across pictures, and producing colored line drawings rather than black and white drawings. The authors state that the words were selected to cover a wide range of frequencies and semantic categories however no statistics are provided on the distribution of the pictures across semantic categories. Upon personal visual examination of the picture set, we note that it covers the following superordinate semantic categories: food/fruit/vegetable, animals (birds, mammals, mollusk, insect), animal body parts, plant, natural element, clothing, house objects, furniture, kitchen utensils, measurement tools, media and communication tool, musical instrument, container/receptacle, building, vehicle/part of a vehicle, sports object, toy/game and desk and writing material (proportions are semantic categories in the selected picture set are detailed in the next chapter). Every picture in the picture set is assigned a name in English by the authors of the study. The names are provided on a separate excel sheet and will not be provided to the committee of experts who will rate and name the pictures. The MultiPic Database can be found on the CogScidotNL (www.cogsci.nl), which is a webpage maintained by Sebastiaan Mothot, Assistant Professor at the department of experimental psychology at the

University of Groningen in the Netherlands. The webpage lists tens of picture sets that are standardized and normed across several variables and that could be used as stimuli in psychological experiments and research. The MultiPic database meets the criteria for an ideal picture set to be used in this study.

To conclude the review of literature, and in light of this long history of challenges and research findings in the MENA region and other areas of the world, there is an obvious need for a picture-naming test that is suitable for Lebanese children. This study is the first of its kind to describe the development of the first Lebanese picture-naming test and the first database of psycholinguistic variables for Lebanese names. The next chapter will describe, in fine details, the methodology and procedures adopted in constructing the Lebanese picture naming test.

Chapter 3: Methods

The study in its entirety was conducted under the approval of the Institutional Board of Research (IRB) of the American University of Beirut. The first phase of the study entails the process of test construction and the development of the first Lebanese psycholinguistic database, while the second phase entails test piloting on a sample of Lebanese children to whom the test is intended. Our theoretical and practical framework is based on the dual-focus approach model for test construction (*Figure 1*). It relies on the contribution of indigenous Lebanese experts in the field of childhood education and development, and it involves the development of test items in several languages simultaneously. The test aims to be culturally and linguistically suitable for Lebanese children in order to reduce the amount of test bias.

Phase 1: Test Construction

The main focus of this study lies in the process of test construction. Test construction is a meticulous process that follows a well-grounded procedure and takes careful consideration of potential biases. The development of the first Lebanese Picture-Naming Test comprised the following stages: (1) selecting the picture set, (2) forming the committee of experts, (3) rating the pictures across psycholinguistic variables, (4) entering and computing the ratings and finally (5) selecting the final pictures to be included in the first draft of the test.

1. Selecting the Picture Set for the Present Study

After extensively reviewing several picture sets, one choice possessed the properties of an ideal picture set in terms of semantic variety, size and picture quality: The MultiPic databank (Dunabeita et al., 2017).




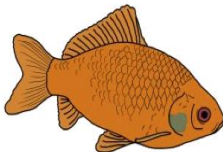

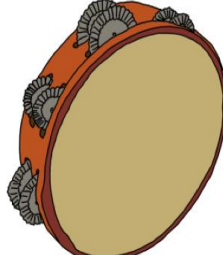
		
Pencil (Desk/Writing Material)	Leaf (Natural Element/ Plant)	Pomegranate (Food)
		
Fish (Animals)	Thumb (Human (Body Part) and Interaction)	Tambourine (Musical Instrument)

Figure 2. Examples of pictures from the MultiPic Databank from various semantic categories

Upon receiving permission from the author of the MultiPic Databank study to use the picture set, the number of pictures (N=750) had to undergo a preliminary reduction in order to prevent fatigue of the committee members during the rating process, avoid the risk of contaminating the quality of ratings by the end of a lengthy task, and match the number of items to other widely used picture-naming tests which may include between 50 to 230 pictures (e.g. BNT, EOWPVT-4, and PPVT-5). To reduce the number of pictures, we decided to adopt a random selection procedure for the following reason: the pictures in the MultiPic set were originally presented in a downloadable folder and listed in a random order (neither listed alphabetically, categorically, nor according to the words' level of difficulty). Therefore, we chose to randomly select every third picture in the picture set in order to avoid introducing selection bias by the researcher. We carried out the method of random selection twice until the pictures were

reduced to almost a little less than a third of the original pool and reached arbitrary number of $N = 219$. Details on the selected pictures and the semantic categories they belong to are reported in the Results section.

2. Selecting the Expert Committee Members

In the original study on the dual-focus approach for test construction, Erkut et al. (1999) mention that, in order to avoid introducing bias into item selection and development, the committee of experts must be indigenous to the target culture and expert on the content of the test. Given that this study's aim is to develop a picture-naming test for children between the ages of 3 and 9 years old, the test requires contribution from professionals who are experts in the field of either early childhood education, language and speech, teaching or special education, test construction or neuropsychological assessments, and, ideally, from different backgrounds.

The committee members were also required to be:

- Native speakers of Lebanese spoken Arabic
- Have lived most of their life in Lebanon
- Familiar with the Lebanese culture
- Additionally, fluent in *either* French or English
- Have a minimum of 7 years of experience working directly with Lebanese children between the ages of 3 and 9 years
- Familiar with the Lebanese school curriculum and have worked in a Lebanese school setting

It was extremely essential that all committee members meet all the above criteria considering that the database of psycholinguistic properties and the picture selection process will heavily rely on their expert judgments. Table 1 shows the list of committee

members and a brief description of their past experiences and credentials. A total of eight professionals took part in the study as members of the committee of experts (mean years of experience = 17). They have all had direct experience in working with children between the ages of 3 and 9 years old and met all the above-required criteria.

Table 1. *Members of the Committee of Experts*

	Current Profession	Highest Level of Education	Years of Experience	Language Proficiency
1	Assistant Director of a Preschool in Beirut	Graduate Studies	>30 years	AR-FR-ENG
2	Doctor in Educational Psychology	Doctorate Studies	>30 years	AR-FR-ENG
3	Speech and Language Therapist	Graduate Studies	>15 years	AR-FR-ENG
4	Special Education Coordinator at a Private School in Beirut	Teaching Diploma in Special Education	7 years	AR-ENG
5	Neuropsychologist	Graduate Studies	>8 years	AR-FR-ENG
6	Consultant on Childhood Education and Literacy Coach	Graduate Studies	>30 years	AR-ENG
7	Child Behavioral Therapist	Graduate Studies	7 years	AR-FR-ENG
8	English Language Homeroom Teacher	Graduate Studies	>15 years	AR-ENG

Note: AR = Arabic. FR= French ENG = English

3. Developing the Materials: Rating Booklets

The purpose of the rating process is to obtain target names for the pictures in Lebanese Arabic, French and English and obtain psycholinguistic measures of cultural familiarity, word frequency and age of acquisition. The rating booklets included the 219 pictures selected randomly by the researcher. The first page of the booklet included a description of the study and explicit instructions to complete the ratings. It also provided definitions of the psycholinguistic variables *cultural familiarity*, *word frequency* and *age of acquisition*, illustrated with examples (Check Appendix A). The booklet was organized in the following format: the pictures were listed in random order

with two pictures sized 10 cm x 10 cm on each page sided by four consecutive columns: (1) Name (in spoken Arabic and in either French or English) (2) Rating for *cultural familiarity* on a five point Likert scale, (3) Rating for *word frequency* on a five point Likert scale, (4) Selection of the *age of acquisition* of the word from the age ranges provided. The researcher completed the first item as an example that could be edited by the committee members.

The researcher printed hard copies of the rating booklet and met with six of the committee members individually, and communicated with two (who were abroad during this period) over the phone. The researcher explicitly and consistently explained the purpose of the study and their role as members of the expert committee. Committee members were specifically asked to:

- 1) Provide a single word that names each picture in Lebanese Spoken Arabic and either French or English or both.
- 2) Rate the cultural familiarity of the object depicted in each picture: *familiarity* was defined as “the extent to which we come in contact with or think about the concept/object/animal in our everyday life”. Committee members were asked to rate the level of cultural familiarity based on how usual or unusual is the picture to a Lebanese child. Given that each member practices in different institutions and schools located in different parts of Beirut, the members were asked to think of children in their realm of work experience when completing the ratings. A 5-point Likert scale was used to rate cultural familiarity where 1 indicates that *the picture is not culturally familiar- a typical Lebanese child will not recognize this picture*, and 5 indicates that *the picture is very familiar to Lebanese children*.

- 3) Rate the word frequency of the object depicted in each picture: *word frequency* was defined as “how often do we come across the word in daily life and in different mediums of language (school textbooks, conversations, written language, media...)”. To highlight the distinction between word familiarity and word frequency, the committee members were provided with two examples “*tarboush or طربوش*” and “*2arweed or قرميد*”, both of which are culturally familiar words or pictures to a Lebanese child but not necessarily often used or encountered. Members who work in a school setting were asked to refer to school textbooks of elementary and preschool children to determine, on average, how frequently encountered is the picture or the word.
- 4) Select the word’s *age of acquisition*: age of acquisition refers to the developmental chronological age when the child learns how to say and use the word expressively. Members were asked to choose the appropriate age range for each picture from nine different ranges (Check Appendix 1).
- 5) Add an observational comment next to pictures they judge to have *low image agreement* (picture is ambiguous or does not resemble the mental image elicited in response to its corresponding word).

All members were informed of the importance of their ratings and were asked to think carefully about their estimation. The members were given about 10 days to fill the rating booklets during which the researcher remained available at all times to answer any question and receive feedback on the rating process. Some of the members called the researcher with specific questions about items and others provided a list of comments and observations in writing. All questions, feedback and observations

provided by the expert committee members regarding the picture set and the rating process are reported in the results section.

4. Processing the Data on the Psycholinguistic Variables

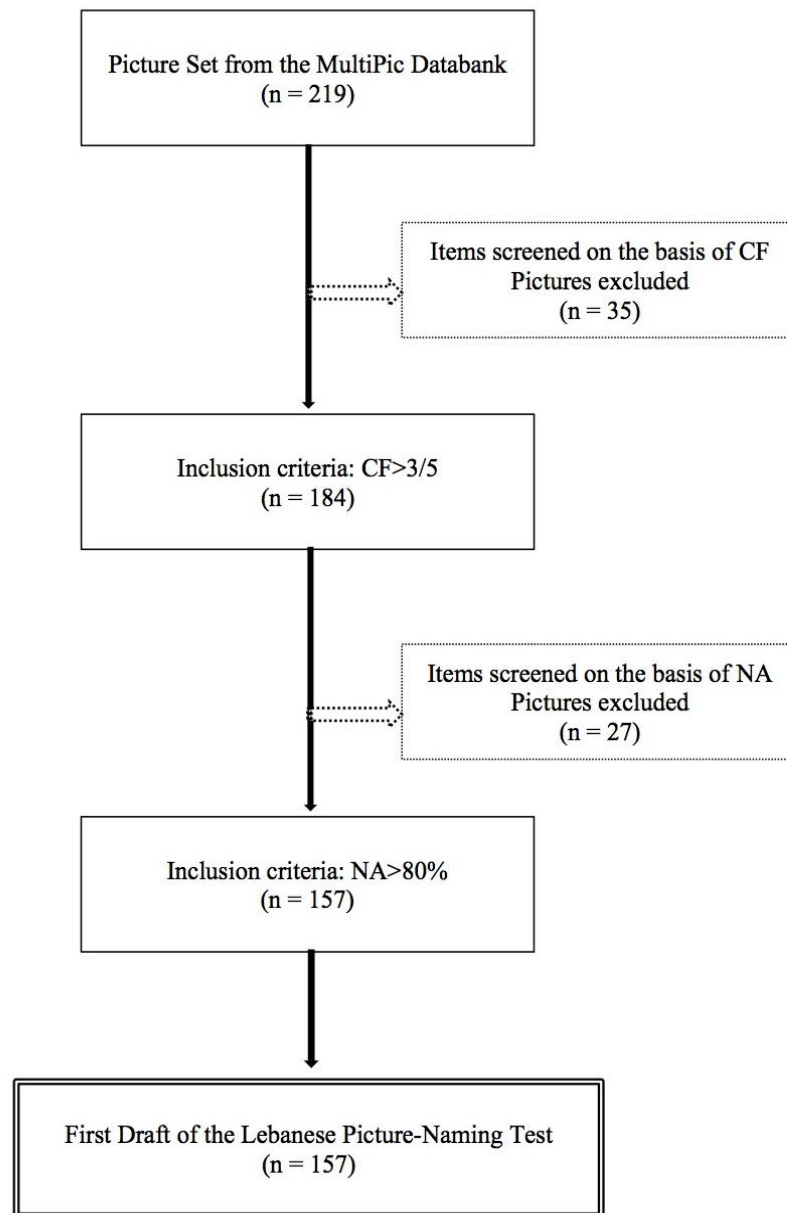
Within approximately two weeks, the researcher received the completed rating booklets from the committee members including the members who were abroad during this phase. All raw data in the rating booklets were entered and computed on Microsoft Excel 2016. Names provided by the committee members were entered verbatim in every language. All committee members consistently provided names in Lebanese Colloquial Arabic (Spoken Lebanese Dialect) except for one member who missed some of the items, five members provided additional names in French and seven members provided additional names in English (three members provided names in all three languages). Cultural familiarity rating and word frequency rating were entered. In the rare cases where the members used decimals for some of the pictures (e.g. cultural familiarity: 3.5/5), the rating was rounded up to the nearest whole number. Following data entry, a descriptive analysis of the ratings was run to obtain an average rating and standard deviation for cultural familiarity, word frequency and age of acquisition for each of the 219 items (reported in the Results section). Because the suggested Age of Acquisition ranges were not all equal (some ranges were 6 months, others 11 months), Age of Acquisition ranges were first transformed into months to allow computing a mean of the members' selected age range for every picture. The means were converted back into years in point decimals where a decimal is a proportion of a year (i.e. 6.5 years is equal to 6 years and 6 months). Frequencies of the picture names provided by the experts were calculated for each language and modal name responses (name with the highest

name agreement among the experts) in Arabic and French and English were selected and assigned to the picture.

5. Compiling the First Draft

After processing the data on the psycholinguistic variables of the pictures, we carried a systematic selection process (illustrated in Figure 3) whereby pictures with a low average rating on cultural familiarity ($< 3/5$) and low percent of name agreement ($< 80\%$) were discarded. None of the remaining pictures had an assigned age of acquisition above 9 years old. The remaining pictures were then compiled in an easel bound test ready to undergo piloting.

Figure 3. Flowchart of the Systematic Process for Picture Selection



Notes. CF = Cultural Familiarity. NA = Name Agreement.

Phase 2: Test Piloting

The purpose of the piloting phase is to evaluate the psychometric properties of the test and compare group performances. Decisions to retain or review pictures will be made accordingly. IRB documents were prepared and approved prior data collection. Permission to access Public Schools in Beirut was granted by the Ministry of Education and Higher Education (MEHE) of Lebanon. Informed consents for parents and school

principals were provided in English, Arabic and French, and child oral assent was provided in English, Arabic and French. The compiled test was reviewed and approved by the IRB for data collection.

Participants

Five out of seven schools and nurseries that received invitations to participate in the pilot study accepted to join. The total pilot sample consisted of 74 Lebanese males and females between the ages of 3 to 9 years. Students were enrolled in 3 private schools and 2 public schools in Beirut's third district. Schools were contacted by the researcher and were conveniently selected based on the principal's approval to take part in the study. Student's inclusion criteria included age between 3 and 9 years and a Lebanese nationality. One student (75th participant) interrupted test administration after item 15 and was excluded from the sample. Students' selection in schools relied on convenience sampling methods stratified across age, gender and type of schooling, as student populations in the selected schools were large and not known by every classroom teacher. Teachers assisted the researcher in selecting students that were easy to reach during break time or before the start of the school day. Nonetheless, efforts were made to maintain equal proportions of males to females, children from all age groups and appropriate proportions of children from public schools to children from private schools (1 public: 3 private). Demographics collected for each participant were age, gender and type of schooling. Some of the public schools in Lebanon group children of different ages in the same grade due to restricted class capacity and student's level of performance, therefore grade level was not an accurate variable to include in the analysis. Only one of the five schools taught exclusively English and Arabic whereas the four other schools taught French, English and Arabic. According to teacher and

administrators' report, none of the students had a pre-existing diagnosis of neurodevelopmental disorders, intellectual disabilities, learning disability and physical disability affecting sensory modalities. Parental consent and child oral assent were received from each participant prior test administration. Table 2 shows the distribution of the pilot sample across age and gender.

Table 2. *Distribution of the Sample across Age and Gender*

Age (years)	Females (N=44)	Males (N=31)	Total (N=74)	
			N	%
3	6	6	12	16.2
4	4	1	5	6.8
5	1	3	4	5.4
6	8	3	11	14.9
7	5	4	9	12.2
8	10	6	16	21.6
9	10	7	17	23.0
Total %	55.4	44.6	74	100.0

Test Administration and Data Collection

Test administration and data collection were carried out exclusively by the student researcher over a period of 4 weeks in school classrooms under standardized testing conditions: noise-free, well lit, containing only the test administrator and participant sitting facing each other. Children younger than 5 years old were at times accompanied by a teacher who remained unobtrusive throughout test administration. Test administrator built rapport with every child by introducing herself, describing the task and receiving oral assent from the child to carry out test administration. The administrator consistently provided the following instructions in English, French or Arabic, depending on the child's preferred language:

I will show you pictures and I will ask you to name each picture using one word.

You can say the name in any language you choose. Just say one word. If you

don't know the name of the picture, we can skip it and try another one! I will be writing down your answers the same way you say them and no answer is a wrong answer. Just say one word to name each picture. Do you have any questions?

Participants were generally compliant during the task and showed enthusiasm by smiling, responding well to verbal praise and sustaining their attention throughout the entire task duration. The test was administered in full to all 74 children. There was no discontinuing rule or ceiling at this point. Participants generally responded promptly to each stimulus and said “pass” or “I don’t know” when they did not have a name for the picture. Average duration of test administration was around 15 min per child.

Spontaneously elicited names provided by students were recorded verbatim in writing during administration and were neither scored as correct or incorrect. No practice items were included at this point since there is no rationale for using one practice item over another, however, the examiner consistently prompted during the first two items if instructions seemed to not be clear (examples of prompts: *what do you call this?* or *This is a...?*). Semantic or phonemic cues were not provided. If the child seemed to be attending to an irrelevant or different aspect of the picture than what is expected, then the administrator provided these types of cues adopted from the EOWPVT-4 (Martin & Brownell, 2011):

- “*What kind?*” was used when the response was too general (e.g. child says fruits for apple).
- “*What else is it called?*” was used when the response was too specific (e.g. child says Mercedes for car).

- “*What is this?*” while pointing at the picture was used when the response described a verb or only a part of the picture was named (e.g. child says flying for helicopter, or, camera for photographer).

Data Entry and Preprocessing

Responses of each participant were entered verbatim on Microsoft Excel 2016. Responses in Spoken Colloquial Arabic were entered in English using the Arabic Chat Alphabet of the “Arabizi” where Arabic words are encoded using Latin script and numbers (For more on Arabizi, refer to Yeghan, 2008). After response entry, responses underwent a thorough and meticulous cleaning process in order to code responses as correct or incorrect. Data preprocessing is described below:

- Long responses were collapsed so that any response that includes the target word and additional details was reduced to only the target word (e.g. for the target word: *hair*, “girl’s hair” or “hair of a girl” were collapsed into *hair*; for the target word *pince* in French, “pince a linge” was collapsed into *pince*).
- Basic variants of the target word in terms of pronunciation due to regional variations in dialect were changed to the assigned target word by the committee members. For example, “mozeh” (a variation of the word “mawzeh” meaning banana) was changed to the target word *mawzeh* and “3alle2a” (a variation of the word “te3li2a” meaning hanger) was changed into *te3li2a*.
- English or French words that are “Arabized” (meaning they were transformed over the years to resemble words in the Arabic dialect) were changed back into their original language. For example, the word “motseekl”, which is the Arabized form of *motorcycle*, was transformed back to *motorcycle*.

- Plural forms of the words were changed to singular. For example “jazar” was changed into *jazra* meaning carrot.

After collapsing and transforming the responses to match the target words, the answers were coded as correct and incorrect. Elicited responses that match the target word verbatim in Lebanese Arabic, English or French were coded as correct. All other responses were coded as incorrect which include responses that are semantically related to the target word but conceptually distinct (“music” for the target word *microphone* = incorrect), responses that are phonetically related to the target word but semantically distinct (“poivre” for the French target word *poire* = incorrect; “fa2as” for the Arabic target word *2afas* = incorrect), responses that are visually related to the target word but semantically distinct (“apple” for the target word *tomato* = incorrect), and responses that are unrelated to the target word. The final score, referred to as “total score”, was tabulated by adding all correct answers.

Data Analysis

Descriptive Statistical Analysis

Data analysis investigated performance across groups, and performance across test items. Although at this point the pilot sample is not large enough to draw conclusions on differences in group performance, some comparison across gender, age, and type of schooling can be carried out to provide us with some information on the properties of the test. Nonetheless, the sample has an acceptable male to female (55.4% of the participants are females) and public school to private school ratio (63.5% of the participants are enrolled in private schools) that is representative of Lebanese students in Beirut, which allows group comparisons to be made to the extent possible. All coded data were entered into an SPSS database version 24. Means and standard deviation of

total score were calculated for age, gender groups and type of schooling groups. To allow for better analysis and group comparison, participants were grouped into three age categories: 3 years to 5 years, 6 years to 7 years and 8 years to 9 years.

Between Groups Analysis. One-way independent analysis of variance is conducted to compare means of test performance across age groups. Two-tailed group comparison is carried between males and females in the total sample, and one-tailed group comparison is carried between children from private schools and children from public schools. Logistic Regression using Enter method is carried with the variables age, gender and type of schooling to check if the model significantly accounts for variance in the total score.

Within Test Items Analysis. At the item level, measures of internal reliability are conducted by calculating Cronbach's alpha in each age group and in the total pilot sample. An item analysis based on the Classical Test Theory (CTT) is conducted across the three age groups to examine inter-item variations. The analysis will evaluate the quality of the items in each age group in order to determine which items to keep, modify or Review. We will first calculate item difficulty index (DIF I; also known as p value), which is defined as the proportion of individuals from the total sample who answer the item correctly. Item difficulty index is calculated with this formula: correct responses on item/total respondents. It ranges between 0 and 1, with higher levels indicating easier items. Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult* = 0.41-0.60, *difficult* = 0.21-0.40 and *very difficult* = 0.00-0.20 (Hetzel, 1997).

Item discrimination (DISC I; also known as d value), which is the ability of an item to discriminate between students of higher and lower abilities, is also calculated.

The total score of participants was entered in descending order and divided into three groups: lowest, middle and highest. The percentage of individuals included in the lowest and highest group may vary. The percentage 27% is usually used because, according to studies, it maximizes the differences in the distribution (Wiersma & Jurs, 1990). The first 27% formed the lower ability group (L) and the last 27% formed the higher ability group (H). Item discrimination index calculates the difference in performance on a particular item between the H group and the L group and is calculated using this equation: $(H \text{ Percent Correct}) - (L \text{ Percent Correct})/100$. Item discrimination index ranges between -1 and 1 with more positive indices indicating higher discriminability power. Item discrimination index is described as *very good* if it's equal or more than 0.40, *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* if it's equal to or below 0.19 (Frisbie & Ebel, 1991). If an item is not able to discriminate between high and low performers across all age groups, then it may indicate that (1) the item's age of acquisition is outside the age range of the sample (3-9), (2) the item may be ambiguous to all ages, or (3) the item is not suitable for the Lebanese children between the ages of 3-9.

Decisions to Retain and Revise Items

The goal of item analysis is to detect which items should be retained and which items should be revised or discarded. Item analysis will adopt a mixed method approach whereby items parameters (quantitative data) along with comments from experts on picture illustrations and drawings (qualitative data) are examined with the aim of eventually eliminating or reviewing items that may be contributing to inaccurate results. An ideal item is one that has moderate difficulty (DIF Index between 0.41 and 0.60) and

high discrimination ($DI \geq 0.40$) (Hingorjo, Jaleel, 2012). However, choosing to discard items based on a cut-off index value will result in losing some items with good content validity therefore some items need to be examined by experts. To illustrate, very easy items such as *dog* and *banana* may have very poor discrimination across all age groups indices because a large percentage of the students get them correct. These items will need to be carefully examined as they may be retained in the test in order to expand the floor of the test during administration. The following decisions will be made based on item parameters:

Decision 1: IF an item has a “very poor” discrimination index ($DISC I < .19$) across the three age groups between 3 and 9 years old, THEN it will be discarded from the picture set UNLESS it is assigned a young age of acquisition that is between 2 and 3 years old.

Items with an age of acquisition of 3 years and younger are expected to have a low level of discrimination since we expect all participants across ages to answer them correctly. Such items may be words like *apple*, *sun*, or *dog* that are assigned by the experts a very young age of acquisition and are therefore likely to show poor discriminatory power across the entire sample. Despite poor item parameters, we will decide to retain these items because they represent good content validity as they are usually children’s first words. These pictures will serve to expand the basal of the test and will be retained in the subsequent draft.

Decision 2: IF an item in the test has good item parameters but was considered by at least one of the experts to have a poor quality of picture illustration or drawing, THEN the item will be examined by two new experts who are blind to the results of the study. The new experts will be provided a rating booklet with the pictures and asked to

rate the picture on (1) typicality, (2) quality, and (3) suggest modifications. If both experts agree that the picture items require modification to its illustration then the item will be revised. A decision tree is included in the results section.

Summarizing the Steps of the Methodology

The development of the picture-naming test adopts the dual-focus approach for test construction in order to develop a test that is culturally and linguistically suitable for Lebanese children. *Figure 4* below shows the steps in the development of our test and how it defers slightly from the classical model of the dual-focus approach. We chose to skip *Step 2* from the classical model (*Step 2*: Operationalize content area to be studied by concepts equally valid in [Arabic, English and French]) because in our case, naming is a rather universal and unidimensional construct that is an elementary process in the use of language (Glaser, 1992). The remaining steps follow the sequence of the classical model with an added outcome to *Step 2* in our model (the rating process), which generated a database of psycholinguistic variables in Lebanese Arabic. The implementation of the dual-focus approach to test development will yield in a first draft of the picture naming test and to ensure minimal bias in the items selected. Additionally, it will allow future revisions and enhancement to the test by repeating *Step 4* and *Step 5* successively to obtain a final draft of the test.

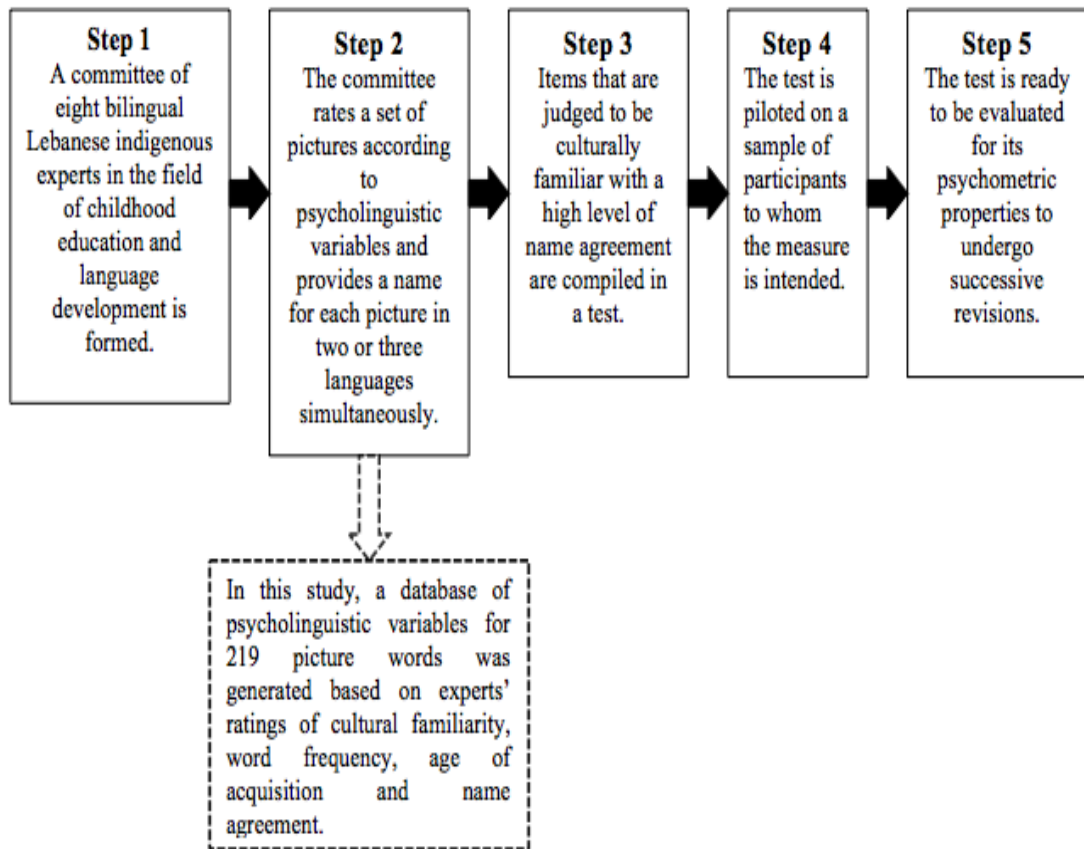


Figure 4. Implementation of the Dual-Focus Approach (Erkut et al., 1999) in the development of the first Lebanese picture-naming test.

Chapter 4: Results

This section reports results from the test development phase and test piloting phase. The first part presents the psycholinguistic database of 219 pictures and semantic categories of the picture words. The second part includes the results from the pilot study and reports group comparisons (across age, gender and type of schooling) and test item functioning.

Outcomes of Test Development

The first phase of the study generated two main outcomes: A Lebanese psycholinguistic database of picture words and the first draft of the picture-naming test.

1. Development of the Psycholinguistic Database for Lebanese Words

The ratings of the eight indigenous Lebanese experts resulted in a Lebanese Database of Psycholinguistic variables for the 219 picture words selected from the MultiPic Databank. For each of the pictures, it provides a name that corresponds to the picture in three languages (Colloquial Lebanese Arabic, English and French), alternative names provided in Colloquial Lebanese Arabic, mean rated word frequency, mean rated cultural familiarity, and mean assigned age of acquisition. The database is found in Appendix B.

Experts' Feedback and Qualitative Observations. During the rating phase, the members of the committee were asked to make qualitative remarks on pictures they consider to have poor quality illustration and could be improved on some aspects. We reported these remarks to the 25 pictures that were found to have good item parameters but were said to be ambiguous. These remarks are included in Appendix C and contributed to the final item selection process. Additionally, the experts wrote to the

researcher their thoughts and reflections on the rating process and some concerns experienced along the way:

1. Committee member #1 (Assistant director of a preschool in Beirut):

“Factors that play a role [in the rating process]: a child might recognize the object in real life but may not be able to name the picture. – The child’s economic background, rich experiences and use of language at home. – Some pictures may be generalized i.e.: tiger instead of cheetah or tree instead of cedars. – Children might recognize the picture but not find the word (i.e. jug). – Age of acquisition might change depending on the child’s upbringing. – Some objects might be very culturally familiar but the picture in the booklet is not easily recognized (i.e. button). – Sometimes, children self-correct when they come across two pictures that are closely related like rooster/hen. Once they see the hen and the rooster one after the other, they’re more likely to name the second picture correctly. The same applies to glove/mitten, thumb/finger. Also, when you draw their attention by prompting with “are you sure”, they self-correct. But this is just an observation. I did not proceed that way. Ratings are typical to the children in the [school where I work]. I reworded the words in both English and French as our kids at [the school where I work] are not very fluent in Arabic, therefore, it is very hard for them to name the pictures in Arabic though they might recognize the picture. – Some pictures are a bit “old style” and not modern graphics similar to what the students are typically encountering nowadays i.e. bread, horse, lion, deer, camel, fingers...”

2. Committee member #2 (Doctor in Educational Psychology):

“Most ratings (familiarity and frequency) were made considering the age specified i.e. relative to the age of acquisition. – Age of acquisition was specified for the acquisition of any correct word that depicts the image. – One thing that was challenging:

determining the age of acquisition by children attending formal preschool vs. those who are not. A child who hasn't been to KG1 would find some animals totally unfamiliar and the word may not be encountered at all in his environment."

3. Committee member #7 (Child Behavioral Therapist):

"I would like to note that there was some confusion in regards to the child familiarity with an image itself versus what it looks like in reality, as well as its technical naming versus the naming in spoken dialect. For example, a child might identify a hiking boot as a shoe, while knowing that its function is for outdoor activities, but his/her naming would be due to instruction/language delivery. That might cause a discrepancy in rating based on the committee members' assumption of the word's acquisition."

2. The First Draft of the Picture-Naming Test

The compiled pictures in the first draft of the picture-naming test were the result of a systematic selection process aimed to discard pictures that are low in cultural familiarity and word agreement. The first selection criterion relied on the picture's average cultural familiarity (CF) rating. Pictures that obtained an average rating on CF of 3.0/5 and higher were included in the first subset. Out of the 219 pictures in the picture pool, 184 pictures met the first selection criteria. The second selection criterion was based on name agreement among the committee members in Lebanese Spoken Arabic. Name agreement among committee members refers to the extent to which they agree on a specific name for the picture. In the process of selecting words based on name agreement, polysemic words (words that could have several correct synonyms) were also eliminated. Modal names were derived for each picture (most frequently provided name). Out of the 184 pictures remaining, 157 pictures were given the same name in Lebanese colloquial Arabic at least 80% of the time across committee

members. Table 3 lists the pictures that were discarded for low average Cultural Familiarity, and Table 4 lists the words that had a low percentage of Naming Agreement among committee members along with alternative names provided by the committee members. All modal responses provided by the members in English matched the names provided by the authors of the MultiPic Database, which meant we were able to keep the assigned picture names in the picture set unedited.

Table 3. List of Discarded Pictures based on Average Cultural Familiarity Rating

Picture Words	Name in Lebanese Arabic*	Average CF
Boomerang	NA	1.17
Boxer	Moulakem	2.38
Broccoli	Broccoli	2.50
Cactus	Sobbeir	2.25
Chain	Janzeer	2.88
Claws	makhlab	1.43
Compass	bikar	2.00
Cone	iqma3 el mourouriah	2.57
Crab	salt3oun	2.75
Dart	sahem	2.38
Deer	ghazel	2.63
Dice	zاهر	2.88
Dominoes	NA	2.75
Dragon	tanneen	2.63
Drums	tabel	2.63
Glove (Mitten)	kfouf	2.71
Greenhouse	khaymeh	2.00
Hippopotamus	7isan el ba7r	2.25
Island	jazira	2.63
Leopard	fahed	2.14
Lumberjack	7attab	1.50
Maze	mata7a	1.50
Megaphone	zammour	2.60
Mermaid	7ourieh	2.13
Microscope	majhar	2.00
Orchestra	fer2a mousi2iyyeh	2.14
Pot (Jug)	jarra	1.63
Puddle	mayy	2.13
Rhino	wa7id el qarn	1.86
Safe	brise (kahraba)	2.83
Saw	menShar	2.88

Saxophone	NA	1.75
Screwdriver	mfak bragheh	2.88
Stamp	tabe3	1.25
Trumpet	bouq	2.63
Wheelbarrow	3arabiyye	2.75

Note. *Modal response provided by the committee members. Responses are written in Arabizi (For more on Arabizi, refer to Yeghan, 2008). NA = absence of provided name in Lebanese colloquial Arabic.

Table 4. *List of Discarded Pictures based on percent Name Agreement*

Picture	Name 1*	Alternative Names	Percentage of Agreement
Grapes	3enab	tout	75.00
Belt	2shat	zennar	71.43
Spider	3ankabout	kertayle	71.43
Steps	daraj	darje	71.43
Beard	da2en	le7iyeh	66.67
Football	kourat qadam	tabeh	66.67
Hug	3abta	ghamra	66.67
Pool	berkeh	masba7	66.67
Runner	3adda2	yarkod	66.67
Singer	moughanne	moutreb/fannan	66.67
Sunflower	douwar el shames	wardeh	66.67
Cap	bornayta	2abbou3a/ta2iyye	60.00
Wasp	na7leh	dabbour	60.00
Baby	Tofol	walad/bobboo	50.00
Baguette	khebez	rgheef/rgheef franje	50.00
Hat	bornayta	2abbou3a/ta2iyye	50.00
Paint brush	fershet	risheh	50.00
Plant	shatleh	zarri3a/nabteh	50.00
Screen	telfez	7asoub/shesheh	50.00
Ship	safineh	bekhra/shakhtoura	50.00
Swimming	sabba7	yasba7/sbe7a	50.00
Teacher	m3allem	estez	50.00
Lamp post	daw shere3	lambet baladiyye/daw	33.33
Shower	NA	NA	0.00
Tractor	NA	NA	0.00
Xbox	NA	NA	0.00

Note. *Modal response provided by the committee members. Responses are written in Arabizi (For more on Arabizi, refer to Yeghan, 2008). Alternative Names = other names suggested. NA = absence of provided name in Lebanese colloquial Arabic.

After discarding items based on low cultural familiarity and low name agreement, 157 items remained. The 157 items were all assigned by the experts an Age

of Acquisition below 9 years old, had a mean word frequency of 3.7/5 and a standard deviation of 0.9, which indicates varied frequency levels across the items.

Variety of Semantic Categories. In order to obtain data on the proportions of different semantic categories present in the picture set, we reached out to the authors of the MultiPic Databank, however, their response was that they had not gone that far in the analysis of the pictures. Therefore, we decided to categorize them according to the following eighteen superordinate categories:

- 3 Natural Kind categories: Animals and Insects, Food, Natural Elements/Plant
- 12 Artifact categories: Furniture, Clothing (or part of) and Accessories, Kitchen Utensils and Appliances, Container/Receptacle, Desk/Writing Material, Vehicle, Tools, Musical Instruments, Media and Communication, Toy/Game, Shapes, Fiction (e.g. dragon).
- 3 Activity categories: Human (or Body Part) and Interaction, Sports, Outdoor Places
- We compared the proportions of semantic categories in the first draft of 157 pictures with the initial set of 219 pictures and found the ratios to be somewhat similar to each other. Table 5 provides a description of the distribution of pictures in both picture pools across semantic categories.

Table 5. *Proportions of Semantic Categories*

Superordinate Category	% in the 157 pictures	% in the 219 pictures
Natural Kind	30.3	29.7
Animals and insects	13.2	12.8
Food	11.2	9.6
Natural element/plant	5.9	7.3
Artifact	58.7	57.1
Clothing (or part of) and accessories	9.9	8.7

Furniture	8.6	7.3
Tools	7.9	8.7
Container/receptacle	6.6	4.6
Kitchen utensils and appliances	5.3	4.6
Vehicle (or part of)	5.3	4.6
Desk/writing material	3.9	4.1
Media and communication tools	3.3	3.2
Toy/game	3.3	5.9
Shape	2.6	1.8
Musical instrument	2.0	2.7
Fiction	0.0	0.9
Activity	11.2	13.3
Human (or body part) and interaction	7.2	7.3
Outdoor places or parts	2.0	2.3
Profession	2.0	3.7

The first phase of the study, which includes the rating process, examination of the psycholinguistic variables and systematic item selection resulted in a final set of 157 pictures that are (1) rated as culturally familiar, (2) have a name agreement above 80% across 8 committee members and (3) cover a diverse selection of semantic categories. The 157 items were compiled in a picture-naming test ready to undergo piloting.

Outcomes of Test Piloting

Pilot Sample Characteristics and Descriptive Results

A description of the pilot group is presented in Tables 6, 7 and 8. The total sample consisted of 74 participants including 33 Males (44.6%) and 41 Females (55.4%). The mean age of the participants was 6.57 years ($SD = 2.13$) ranging between 3 years and 9 years. Forty-seven (47) students attended private schools (63.5%) and 27 students attended public schools (36.5%).

Table 6. *Distribution of the Sample Across Gender and Type of Schooling Separately*

		N	%
Gender	Male	33	44.6%
	Female	41	55.4%
Schooling	Private	47	63.5%
	Public	27	36.5%

Table 7. *Descriptive of the Sample Characteristics Age and Total Score*

Demographics	N	Minimum	Maximum	Mean	SD
Age	74	3.00	9.00	6.57	2.14
Total Score	74	46.00	142.00	103.47	23.70

The sample was divided into three groups according to age ranges (Table 8). Participants between 3 years and 5 years old (N = 21) received a mean total score of 80.52 ($SD = 23.32$) on the picture naming test, participants between 6-7 years old (N = 20) received a mean total score of 110.15 ($SD = 18.9$) and participants between 8 and 9 years old (N = 33) received a mean total score of 114.03 ($SD = 15.44$). Mean total score of all participants was 103.47 ($SD = 23.69$) with scores ranging between 46 and 142. Males (N=33) received a mean total score of 98.21 ($SD = 22.01$) whereas females (N=41) received a mean total score of 107.70 ($SD = 24.41$). Table 9 shows mean performance of participants stratified by age group, gender and type of schooling.

Table 8. *Mean Performance Across Age Groups*

Age	N	Mean	SD	SE	Min	Max
3 to 5	21	80.52	23.32	5.08	46.00	119.00
6 to 7	20	110.15	18.99	4.24	70.00	142.00
8 to 9	33	114.03	15.44	2.68	68.00	138.00
Total	74	103.47	23.69	2.75	46.00	142.00

Table 9. *Mean Performance Stratified by Age Group, Gender and Type of Schooling*

Age Groups	Gender	Schooling	M	N	SD
------------	--------	-----------	---	---	----

		Private	76.11	9	19.47
	Male	Public	87.50	4	24.03
		Total	79.61	13	20.66
3 to 5	Female	Private	83.57	7	30.51
		Public	71.00	1	-
		Total	82.00	8	28.60
Total	Total	Private	79.37	16	24.27
		Public	84.20	5	22.08
		Total	80.52	21	23.32
		Private	105.83	6	14.57
	Male	Public	82.00	1	-
		Total	102.42	7	16.07
6 to 7	Female	Private	123.77	9	10.47
		Public	93.00	4	19.71
		Total	114.30	13	19.72
Total	Total	Private	116.60	15	14.87
		Public	90.80	5	17.76
		Total	110.15	20	18.99
		Private	115.00	8	8.12
	Male	Public	113.80	5	8.46
		Total	114.53	13	7.92
8 to 9	Female	Private	129.25	8	9.91
		Public	103.33	12	16.40
		Total	113.70	20	19.01
Total	Total	Private	122.12	16	11.43
		Public	106.41	17	15.07
		Total	114.03	33	15.44

Preliminary Analysis: Exploring Assumptions

A missing data and mis-entered data analysis was conducted, and the data were screened for univariate outliers [standardized z -scores larger than 3.29 ($p < .001$, two-tailed) were used as criteria for univariate outliers]. The data contained no mis-entered

data, missing data or univariate outliers. Regarding assumptions of normality, we expect that normality of total scores would not be necessarily met across our independent variables, given that the participants represent a pilot sample. In fact, the assumption of normality was not met for the variables total score for all participants $W(74) = 0.94, p = .001$, age $W(74) = 0.87, p < .001$, total score of males $W(33) = 0.906, p = .007$, total score of females $W(41) = 0.92, p = .009$, and total score of children in private schools, $W(47) = 0.96, p = .001$. On the other hand, total score of children from public schools met the assumption of normality with $W(27) = 0.956, p = .29, ns$, however, because the rest of score distributions deviate from normality, we were restricted to the use non-parametric tests. Data transformation was not recommended due to the small sample size and selection criteria. Transforming the data helps as often as it hinders the accuracy of the analysis (Games, 1984).

Data Analysis at the Sample Level

1. Comparing Means across Age Group

We conducted a one-way ANOVA to compare mean test score across the three age groups. Given that the data violates the assumption of homogeneity of variance, Welch's F-test are reported. The results in Table 10 indicate that there was a significant difference between mean scores across the three different age groups: $F(2, 37.87) = 16.89, p < .001$.

Table 10. *Robust Tests of Equality of Means*

Total Score	Statistic ^a	df1	df2	Sig.
Welch	16.895	2	37.874	.000

a. Asymptotically F distributed.

Games-Howell post-hoc tests were used. These tests revealed significant differences between the first age group (3-5) and both the second (6-7), $p < .001$, and third age group (8-9), $p < .001$, however, there was no significant difference between the second group and the third group, $p = .723$, *ns*.

Table 11. *Multiple Comparisons of Total Score using Games-Howell*

(I) Age Groups	(J) Age Groups	Mean Difference (I-J)	SE	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
3 to 5	6 to 7	-29.62*	6.62	.000	-45.79	-13.45
	8 to 9	-33.50*	5.75	.000	-47.66	-19.34
6 to 7	3 to 5	29.62*	6.62	.000	13.45	45.79
	8 to 9	-3.88	5.02	.723	-16.19	8.43
8 to 9	3 to 5	33.50*	5.75	.000	19.34	47.66
	6 to 7	3.88	5.02	.723	-8.43	16.19

*. The mean difference is significant at the 0.05 level.

2. Comparing Means across Gender and Type of Schooling

A Mann-Whitney-*U*-test was carried out to compare performance of female participants ($N = 41$, $M = 107.71$, $SD = 24.41$) to male participants ($N = 33$, $M = 98.21$, $SD = 22.02$) across all age groups. Results in Table 12 show that females scored significantly higher than males on the Lebanese picture naming test with $U = 492.50$, $z = -2.00$, $p = .045$ and a small-to-medium effect size $r = .23$. Additionally, Table 13 shows that the total number of students enrolled in private schools ($N = 47$, $M = 105.80$, $SD = 25.99$) scored significantly higher than the total number of children enrolled in public schools ($N = 27$, $M = 99.40$, $SD = 18.81$) with $U = 478.5$, $z = -1.75$, $p = .040$ (*one-tailed*) and a small-to-medium effect size of $r = .20$.

Table 12. *Mann-Whitney U test- Gender Differences across Total Score*

	Total Score
Mann-Whitney <i>U</i>	492.50

Wilcoxon W	1053.50
Z	-2.00
Significance (two-tailed)	.045

Table 13. *Mann-Whitney U test- Schooling Differences across Total Score*

	Total Score
Mann-Whitney U	478.50
Wilcoxon W	856.50
Z	-1.75
Significance (one-tailed)	.040

3. Predictive Analysis

Next, we conducted a linear regression analysis predicting total score using age, gender and type of schooling as predictors. Table 13 shows that the model accounts for 49% of the variance in total score at the sample level ($R^2 = .49$, $F(3, 70) = 22.74$, $p < .001$). The adjusted R-Square (.47) shows little shrinkage from its unadjusted value (.49), which indicates that the model would account for about 2% less variance in the population. Table 14 shows that the variable Age ($b = 7.80$, $\beta = .70$, $t(70) = 7.70$, $p < .001$) significantly predicted total correct responses. The beta value indicates that as age increases, total score increases. Similarly, type of schooling ($b = -18.76$, $\beta = -.38$, $t(70) = 4.24$, $p < .001$) significantly predicted total correct responses. The beta value indicates if a student belongs to a private school, they're more likely to receive higher total score on the test. On the other hand, gender did not significantly predict total correct responses ($b = 5.73$, $\beta = .12$, $t(70) = 1.4$, $p = .17 ns$).

Table 13. *R, R Square, Adjusted R Square*

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				Durbin-Watson	
					R Square Change	F Change	df1	df2		Sig. F Change
1	.70	.494	.472	17.22	.494	22.74	3	70	.000	1.66

Table 14. *Regression Parameters*

Model		B	SE B	β
1	(Constant)	68.95	9.06	
	Gender	5.73	4.10	.12
	Age	7.80	1.01	.70***
	Schooling	-18.76	4.43	-.38***

***. Correlation is significant at the 0.001 level.

Data Analysis at the Item Level

First, Cronbach's alpha coefficients were computed by age groups (3-5; 6-7; 8-9) and for the total sample (N= 74). These coefficients shown in Table 15 are high ranging from 0.90 to 0.95 for the various age groups, which indicates a large scale homogeneity of the test items.

Table 15. *Internal Reliability Measures*

Age Groups	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
3 to 5	.958	.955	149
6 to 7	.935	.934	143
8 to 9	.909	.910	141
Total (N=74)	.958	.957	155

Item analysis examined participant responses to individual test items in order to assess the soundness of the items and how they are functioning in the test as a whole. Item Difficulty Index (DIF I) and Item Discrimination Index (DISC I) were calculated for each item in the test. Items were grouped according to their level of difficulty and discrimination pertaining to each of the three age groups (Tables 16, 17, 18). Darker boxes signify poor item parameters. These include items that have a poor discrimination index.

Appendix C shows item parameters of all 157 pictures across the three age groups along with qualitative comments from the experts. Qualitative data is also integrated in the decision making process. Pictures that were considered by an expert to

have an ambiguous or unclear illustration were examined by two new experts who are blind to the item parameters and the previous ratings. If both new experts agree that at least one aspect of the illustration needs revision then the item will be set aside to be revised and edited.

Table 16. *Matrix of Test Item Parameters: Item Difficulty Index x Item Discrimination Index (Ages: 3-5)*

DISC I DIF I	Very Good (0.40<)	Reasonably Good (0.30-0.39)	Marginal Item (0.20-0.29)	Poor (<0.19)
Very Easy (0.81<)		Tree (Round) Tree Knife Strawberry Rain Bird Tomato Moon Giraffe Orange Duck Table Fork Sofa Gift		Car Dog Apple Balloon Banana Shoe Flower Fish Glasses Butterfly Hand Carrot Sun Cow Ice Cream Telephone Scissors Foot Spoon Fire Pencil Star Watch Horse Square
Easy (0.61-0.80)	Heart Stairs Hair Pear Bag Guitar Tiger Chalkboard Nose Finger Lamp (Bulb) Book Candle Piano Corn Helicopter Dolphin Hamburger Pineapple			Umbrella Snail Salt Mouse Key Motorcycle Lion Basket Leaf
Moderately Difficult (0.41-0.60)	Train Puzzle Wheel Fan Road Cage Bicycle Ant Fruits Thumb Glove Hair Brush Bell Face Necklace Glass Pot	Pacifier Bottle Box Peg Teacher (Female)		Chicken Suitcase Bin Rooster
Difficult (0.21-0.40)	Rectangle Lamp Lemon Mushroom Pomegranate Bone Bench Comb Cheese Cupcake Hammer Broom Parachute Shovel Microphone Logs Skateboard Hanger Button Computer	Parrot Camera Rose Goat		Presto Taxi Notebook Boot
Very Difficult (>0.20)	Feather Jar Goal Diamond T-Shirt Stapler Envelope	Milk Zipper Stool Trophy Tambourine Camel Artist Stadium Mug Lock		Needle Tunnel Chair Scale Torch Fire Extinguisher Goalkeeper Keyboard Photographer Briefcase Teapot Skeleton Fish Tank Brain Tray Bird (Pigeon) Tie Lace

Note. Words with poor DISC I (<.19) to be examined and possibly revised

Table 17. *Matrix of Test Item Parameters: Item Difficulty Index x Item Discrimination Index (Age: 6-7)*

DISC I DIF I	Very Good (0.40<)	Reasonably Good (0.30-0.39)	Marginal Item (0.20-0.29)	Poor (<0.19)
Very Easy (0.81<)	Balloon Giraffe Bone Carrot Duck Fruits Ant		Nose Table Tree (Round) Ice Cream Shoe Moon Tree Horse Strawberry Lion Hamburger Scissors Watch Tomato Glasses Chalkboard Bag Stairs Rain	Dog Hand Car Bird Fish Banana Sun Heart Butterfly Flower Orange Star Foot Key Cow Pencil Apple Fire Knife Umbrella Necklace Gift Hair Telephone Bicycle Candle Motorcycle
Easy (0.61-0.80)	Fork Finger Corn Salt Bell Basket Square Pineapple Goat Guitar Leaf Microphone Pear Cage Zipper Broom Lemon Tiger Camera Mushroom Puzzle Snail Road Bottle Jar Dolphin Face Lamp Teacher (Female) Piano Button Hair Brush Artist Hammer Skateboard Skeleton		Sofa Feather Book Fan Trophy Train Box Wheel Suitcase	Spoon Mouse Stapler Rose Chicken
Moderately Difficult (0.41-0.60)	Helicopter Computer Hanger Thumb Brain Glove Rooster Cheese Lock T-Shirt Shovel Mug Diamond Fish Tank Rectangle Cupcake Tray Camel Tie Bench Pacifier Tunnel		Logs Pot Needle Glass Parrot Notebook Teapot Comb	Lamp (Bulb) Taxi Pomegranate
Difficult (0.21-0.40)	Torch Goalkeeper Tambourine Envelope Chair		Boot Stadium Parachute Stool Lace Keyboard	Peg Goal Briefcase Presto Milk Bin Scale
Very Difficult (>0.20)			Fire Extinguisher Photographer	Bird (Pigeon)

Note. Words with poor DISC I (<.19) to be examined and possibly revised

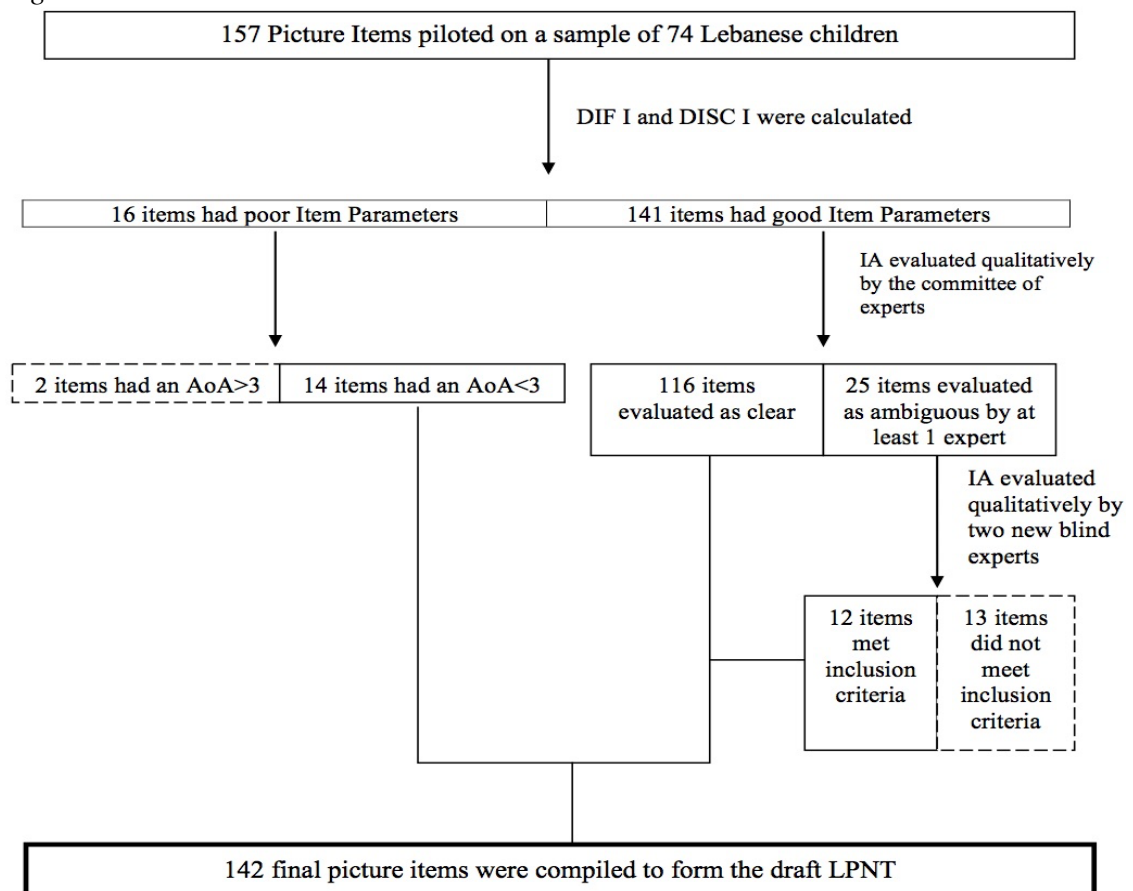
Table 18. *Matrix of Test Item Parameters: Item Difficulty Index x Item Discrimination Index (Ages: 8-9)*

DISC I DIF I	Very Good (0.40<)	Reasonably Good (0.30-0.39)	Marginal Item (0.20-0.29)	Poor (<0.19)
Very Easy (0.81<)	Telephone Giraffe Brain Cage Piano Microphone	Finger Leaf Candle	Bird Corn Fire Scissors Bone Feather Ant Pineapple Train Rose Spoon Artist Camera	Dog Hand Apple Car Nose Banana Sun Tree (Round) Ice Cream Heart Butterfly Tree Lion Star Key Stairs Table Glasses Fish Tomato Hair Flower Knife Strawberry Chalkboard Pencil Hamburger Watch Motorcycle Balloon Carrot Cow Moon Umbrella Fork Salt Orange Bicycle Foot Bag Shoe Mouse Fruits Bell Face Stapler Horse Lamp (Bulb) Sofa Duck Tiger Basket Wheel Gift Fan
Easy (0.61-0.80)	Trophy Pear Zipper Dolphin Mushroom Button Computer Parrot Snail Diamond Box Square Puzzle Helicopter Goat Glove Skeleton	Guitar	Chicken Camel Broom Lock Bottle Lamp	Necklace Pot Book Teacher (Female) Pomegranate Suitcase Road Bin Rain Hanger Glass
Moderately Difficult (0.41-0.60)	Thumb Tie Mug Skateboard Rectangle Lemon Rooster Goal Parachute Goalkeeper Hammer Hair Brush Jar Fish Tank Shovel	Cheese Needle Teapot Milk Scale	Tray	Comb Taxi Presto Stadium Logs Briefcase Boot Peg Envelope
Difficult (0.21-0.40)	Cupcake T-Shirt Torch Keyboard Notebook Tunnel Bench		Pacifier Chair Lace	Stool
Very Difficult (>0.20)	Tambourine		Bird (Pigeon) Photographer	Fire Extinguisher

Note. Words with poor DISC I (<.19) to be examined and possibly revised

Figure 5 describes the decision tree for reviewing or including items. Of the 16 items that need to be examined due to poor item parameters across the three age groups (DISC I < .19), 14 items were assigned an Age of Acquisition of 3 years and below and were therefore retained in the picture set in order to expand the basal of the test. Of the 141 items that had good item parameters, 25 items were considered by the experts to have flaws in their illustrations and were consequently evaluated by two new blind experts. The qualitative comments on the picture illustration are reported verbatim in Table 19. Of the 25 pictures, 13 pictures were agreed upon by both new raters to require editing or re-illustration (did not meet inclusion criteria to be compiled in the final draft). These items will be set aside and re-illustrated before conducting future studies.

Figure 5. Decision Tree to Select the Final Items in the LPNT



Note: DIF I = Difficulty Index. DISC I = Discrimination Index. AoA = Age of Acquisition. IA = Image Agreement. LPNT = Lebanese Picture-Naming Test.

Appendix D includes the table listing all the items that were examined and the decision to either retain them or revise them after expert examination. The table also reports suggestions from the experts to render the illustration of the item clearer.

Chapter 5: Discussion

The purpose of this study was to construct and pilot the first Lebanese picture-naming test that is culturally and linguistically suitable for Lebanese children. The last section will summarize the outcomes of the study that consist of (1) the development of the first Lebanese Psycholinguistic Database, (2) the draft Lebanese Picture-Naming Test, (3) comparison of test performance across group variables (age, gender and type of schools) and how they relate to existing literature and finally (4) performance across test items. We will conclude the study by listing limitations and future directions.

Summarizing the Outcomes of the Study

A Psycholinguistic Database for Lebanese Words

Our first aim was to provide researchers and clinicians with a database of psycholinguistic variables pertaining to 219 picture words based on the ratings of eight indigenous Lebanese experts in the field of childhood education and development (Appendix B). The database includes measures of word frequency, cultural familiarity, age of acquisition, names in Colloquial Lebanese Arabic, French and English, name agreement, and alternative responses provided.

A psycholinguistic database can provide researchers and practitioners with a set of picture words with known semantic and psycholinguistic variables. It is for use in the domains of speech and language therapy, childhood education and special education, educational psychology, psycholinguistics, neuropsychology and psychopathology. This database can be added to other existing psycholinguistic databases in the Arab region (Alyahya & Druks 2015; Bakhitar, Nilipour & Weekes, 2013; Boukadi, Cirina & Wilson, 2015; Ghasisin, Yadegari, Rahgozar, Nazari & Rastegarianzade, 2015;

Khwaileh, Body & Herbert, 2014), should potentially be expanded, and enlarged in future studies.

Introducing the Lebanese Picture-Naming Test

The second aim of the study was to use the measures of psycholinguistic variables derived from the ratings to develop the first draft picture-naming test while implementing the dual-focus approach for test development. Pictures with a high rating on cultural familiarity and name agreement were compiled in a test while the rest of the pictures were discarded. The LPNT improves upon currently used imported tests in that test items were chosen based on (1) cultural familiarity, (2) name agreement, (3) good item parameters across three age groups, and (4) low ambiguity. The selected pictures had different frequencies of use in daily life and spanned a wide array of semantic categories, more specifically, 18 superordinate categories that fall under “Natural Kind” (e.g. plants, animals, food), “Artifact” (e.g. kitchen utensils, toys, furniture) and “Activity” (e.g. human body parts, sports). According to the literature, children first identify pictures at the basic level of abstraction, followed by *superordinate* and *subordinate levels* (Brownell, 1978; Hutcheon, 1970; Smith, Balzano, & Walker, 1978). For example, an armchair (subordinate) is a type of chair (basic), which is a type of furniture (superordinate) (Mervis & Crisafi, 1982). Therefore, the variety of semantic categories at the superordinate level adds to the construct validity of the test in that it evaluates knowledge of words that span several vocabulary categories and are not restricted to a few.

Comparison of Test Performance across Group Variables

The dearth of locally developed picture-naming tests or measures of vocabulary that are culturally and linguistically suitable for Lebanese children makes it difficult to

compare our study's results to previous local findings. However, we could draw some comparisons to previous literature on the effect of gender, schooling and age on gain in vocabulary. Although the purpose of the pilot study does not focus on drawing conclusions on group performance, results derived from the pilot study seem to be consistent with previous literature. We do aim however, to examine sources of variability in vocabulary in Lebanese children in addition to associations of vocabulary to other cognitive factors in future norming studies using the LPNT.

Age as a Contributing Variable. Preliminary results from test performance comparisons across age groups showed that students between the ages of 3 to 5 years performed significantly lower than students aged 6 to 7 and 8 to 9, whereas there was no significant difference in test performance between children aged 6 to 7 and 8 to 9. Many studies attest to the fact that age is a major contributing factor in vocabulary gain (Basilio, Puccini, Silva, & Pedromónico, 2005; Bates, Dale, & Thal, 1995; Vogt, Douglas, & Aussems, 2015). This was quiet evident in our study with older children performing better on the picture-naming test and age being the highest contributing factor to increase in total score. However, future studies using the LPNT should further investigate why performance between the two upper age groups 6-7 and 8-9 was not significant. Possible explanations could be low levels of difficulty in the items and assigned age of acquisition being less than 9 years old. Future studies may want to flag items that were assigned an age of acquisition above 5 years old but that had low levels of item difficulty across both age groups 6-7 and 8-9.

Gender as a Contributing Variable. Gender is another investigated factor in children's vocabulary growth. Although gender differences are not always evident on vocabulary test performances (Bates et al., 1988; Gottfriend & Bathurst, 1983; Zec,

Burkettm, Markwell & Larsen, 2007), some studies show that females perform better than males (Randolph et al., 1999). This is also consistent with our study's results that report better performance of females than males however, when gender was added to the model along with age and type of school, it did not show to be a significant contributing factor to vocabulary gain.

Type of School as a Contributing Variable. Results also show that children enrolled in private schools in this study performed significantly higher than children enrolled in public schools. Although there are no reported numbers by the government or researchers that show evidence of a difference in SES between children from private schools compared to children from public schools, reports from Lebanon indicate that children enrolled in public schools receive less funding from the government, have lower rates of success on intermediary examinations, receive education from less qualified teachers and personnel (Frayha, 2009; PNUD Report, 2009). Therefore, we assume that those enrolled in private schools generally have the financial means to afford this type of education and are more likely to belong to a middle or upper social class. Studies on the relationship between socio-economic background and vocabulary show that children from lower SES build their vocabularies at a slower rate than children from high SES (Dollaghan et al.; Feldman et al., 2000; Rescorla & Alley, 2001). Consistent with this are findings in our studies showing that type of school is a contributing factor in vocabulary with children in private schools performing better than children in public schools while controlling for age and gender. This could be due to the reasons mentioned above that relate to the quality of education received at schools in addition to less opportunities for stimulation, learning in the home environment and complexity of language parents use with their children.

Predictive analysis revealed that age was the strongest contributing factor to increase in test score, followed by type of schooling, with gender showing no predictive power.

Analysis at the Item Level

Analysis at the item level showed that the 157 test items had an excellent internal consistency across the three age groups and in the total sample indicating a high homogeneity between the test items. Item parameters were calculated to detect items with poor functioning. A decision tree was implemented whereby items with poor item discrimination across the three age groups were retained *only* if they were assigned an AoA below 3 years and items evaluated as having poor image agreement were retained *only* if both new blind experts agreed to include them in the final selection of items.

Item analysis, which included both quantitative and qualitative data, resulted in reviewing the quality and illustration of 15 items and preserving 142 items. These items were compiled to form what we can call the draft Lebanese Picture Naming Test (LPNT). The LPNT is the result of a rigorous selection process that included a committee of experts to evaluate cultural familiarity and name agreement, item piloting, and evaluating item functioning. Additionally, the test was found to have homogeneous items for measuring naming ability (Cronbach's alpha ranging from 0.90 to 0.95 across the three age groups and in the total sample).

Test Development Procedure and Best Practice Methods

Throughout the process of test development, we consciously made decisions to reduce test and item bias by referring to ITC guidelines for test development and the implementation of the dual-focus approach method. According to the ITC guidelines for large-scale assessment of linguistically diverse population (ITC, 2018), developing a

test for several languages should consider that the items are equally familiar for the sample target population. We also avoided reference to historic contexts or references that are culture-specific and included experts that are knowledgeable of the target culture and proficient in the test's target languages. We attempted to form a pilot group that includes test takers from all linguistic and sub-cultural groups.

We adopted the most suitable method for test development for a Lebanese sample, the dual-focus approach, to allow the development of test items in several languages simultaneously. The dual-focus approach has been adopted in the development of tests and questionnaires (e.g. Kummervold et al., 2008; Language Surveys; Potaka, & Cochrane, 2004; Tropp, Erkut, García Coll, Alarcón, & Vázquez García, 1999). The two main features of the dual-focus approach were successfully implemented in our study: horizontal collaboration of indigenous experts during item selection, and adoption of a concept-driven approach through the involvement of experts that are knowledgeable in the content area and purpose of the test. This process adds to the test adaptation in that it involves the creation of items from scratch and assigning a culturally familiar name to the item in all languages used.

Limitations

Several limitations related to the choice of stimuli, rating process and item selection emerged during the implementation of the study. The initial random picture selection by the researcher to decrease the number of pictures (from 750 to 219) may have excluded some pictures that would have shown good item functioning in the piloting phase or further increased semantic variety. The committee of experts formed a modest-sized group of 8 individuals, which may have influenced the rating process. Measuring name agreement across the raters was a tricky step since responses were

provided in three languages. Ideally, name agreement is estimate by calculating an *H value*, which provides detailed information on the distribution of names across the committee members. However, given that the names are provided in three different languages, H value may show inaccurate results that point to high variation in names and therefore no derivation of target names. Because responses were allowed in French, English or Arabic, therefore, H values of the pictures would have been high and would not be able to detect name agreement if we get 3 modal responses each in one language. The alternative was to calculate name agreement manually for each of the three languages. This sometimes resulted in very small differences between the modal response and second most frequently provided response. Another limitation relates to the Lebanese dialect and the regional variations of names in terms of pronunciations, plural forms, and use. For example, one-word *spoon* had at least three different responses provided in Arabic (“*ma3l2a*”, “*mal3a2a*”, and “*mal3aqa*”). Reducing the three responses into one of the provided answers dismissed several other variations when administering the test in different areas in Lebanon. This highlights the importance of having the test administrator familiar with the Lebanese culture and dialect in order to different name variations from mispronunciations or sound deletion, substitution or additions. Efforts were also made to render the pilot sample as representative as possible of the Lebanese population however, this was easier to achieve on the type of schooling level, and more difficult across age ranges and gender. It would have been ideal to have both genders equally represented in the pilot sample. And finally, going back to the theoretical framework of the study, choosing to assemble a test rather than adapt an already existing one has implications for the level of equivalence. When we chose to assemble the test, we also prevented it from being used

across-cultures to obtain score comparison. In other words, this is a test that was assembled to suit Lebanese children and may not be handy in comparing naming ability across cultures.

Future Directions

This study is the first step in the development of the LPNT. The test is expected to undergo validity and reliability studies followed by norming studies on a large sample that is representative of children in Lebanon. We could foresee some revisions that will take place before carrying out the norming study.

Test Format. Pictures will be ordered according to item difficulty. The order may undergo modification after carrying out norming studies. The test will maintain the same design: easel bound, with large colored pictures presented one at a time.

Administration. After collecting norms, we can determine the basal (test entry level) and ceiling (test termination level) by calculating how many students from a certain age stop answering correctly after a certain number of false consecutive answers.

Test Content. The test now contains items that are judged to be culturally familiar, with a high name agreement, from various semantic categories, good item parameters and non-ambiguous. Items that were revised by the two blind experts will be re-illustrated while taking into consideration their suggestions.

Scoring. One aspect to be reviewed is the scoring system of the responses. Studies on naming performance describe different types of errors that may be indicative of specific difficulties. Dell et al. (1997) classify types of errors according to their relationship to the target response: (1) semantic error (“apple” for “orange”), (2) formal or phonological error (“hair” for “chair”), (3) mixed error that is both semantically and phonologically related to the target word (“rat” for “cat”) or (4) an error that does not fit

with any of the former categories. Some errors may be diagnostically. Currently, responses can be coded as correct (modal response provided by the experts) or incorrect. We may however consider adding query to some of the responses. If the participant provides a word that is synonymous to the target word (listed in the scoring sheet), the examiner should cue with “give me a better word”. No partial scores will be provided.

Normative Studies. The normative sample will be large enough to represent the population of children in Lebanon between the ages of 3 and 9 years. At this point, age range will not go beyond the age range of the pilot sample unless test content is modified and additional pilot studies are carried out. Given that age, gender and type of schooling were found to have an impact on test performance, these demographic variables will be included in our normative study. The normative sample will be stratified according to area of residency (governorate or *muhafazah*), age (in years and months), gender (males; females) and type of schooling (private; public). Raw scores on the naming test will be transformed to standard scores and will be used as the dependent variables in the validity and reliability studies.

Psychometric Properties. Properties of the test and test items reported in this study should be regarded as preliminary. Validity and reliability studies will be carried out. The performance of a control group and a group of children with an established diagnosis of a language disorder will be compared. The groups will be matched across demographics. Next, if we plan to increase diagnostic validity of the test, we will then evaluate the diagnostic accuracy of the derived scores from the group of children with a language disorder to calculate sensitivity, specificity and efficiency (classification

accuracy of the test), predictive value of a positive test and predictive value of a negative test.

Stability of the test will be assessed using test-retest method on individuals who were not part of the norming sample. Reliability studies will be carried out under the same standardized conditions as the pilot study by trained test administrators.

If we plan to exclude type of schooling from the sample stratification, then efforts should be made to create a measure that is free from association to type of schooling. Differential Item Functioning (DIF) analysis can be completed to detect differential item functioning between groups enrolled in private and public schools. Items that show a high DIF between across two types of schooling should be carefully examined. Additional analysis that go beyond the scope of this study could be conducted such as the effects of language proficiency, dominant language, being bilingual or trilingual on the type of naming responses and overall score.

Test Use. Currently, the test can be used for screening purposes. The score generated will provide the examiner with information on the individual's level of expressive vocabulary compared to a normative sample. Diagnostic decisions cannot rely solely on the score of the LPNT but require complementation of a larger battery of tests and clinical corroboration. However, if the LPNT develops diagnostic items or enhances its specificity and sensitivity to clinical populations (individuals with language disorders and aphasia) then it may provide the clinician with more information on the examinee's ability.

Conclusion

Currently used tools to measure naming abilities in Lebanese children present major limitations, and the need of a naming test that is culturally appropriate to the

Lebanese dialect is indispensable. Given that naming ability is important for the evaluation of language, reading, neurodevelopmental and neurocognitive abilities, it is necessary to have access to a test that can be used confidently by those who work in the field of psychology, education, language and measurement in order to guide their decisions, and plan better interventions.

To conclude, the draft LPNT is the first picture-naming test suitable for Lebanese children between the ages of 3-9 years that includes items that are culturally familiar to Lebanese children, with a high level of name agreement, good item parameters and non-ambiguous. Our hope is that results from this preliminary study can be used in future normative studies.

References

- Adlington, R. L., Laws, K. R., & Gale, T. M. (2009). Visual processing in Alzheimer's disease: Surface detail and colour fail to aid object identification. *Neuropsychologia*, *47*, 2574–2583.
- Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods*, *31*(3), 531-552.
- Kaufman, A. S. (2018). Contemporary intellectual assessment: Theories, tests, and issues. Guilford Publications.
- Alyahya, R. S., & Druks, J. (2016). The adaptation of the Object and Action Naming Battery into Saudi Arabic. *Aphasiology*, *30*(4), 463-482.
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Pub.
- American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073. Available from the APA Web site: <http://www.apa.org/ethics/code2002.html>
- Araújo, S., Reis, A., Petersson, K. M., & Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology*, *107*(3), 868.
- Ardila, A. (2007). Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology*, *22*(3), 297-307.

- Bacha, N. N., & Bahous, R. (2011). Foreign language education in Lebanon: A context of cultural and curricular complexities. *Journal of Language Teaching and Research*, 2(6), 1320.
- Bakhtiar, M., Nilipour, R., & Weekes, B. S. (2013). Predictors of timed picture naming in Persian. *Behavior Research Methods*, 45(3), 834-841.
- Bashur, M. (2004). *Higher education in the Arab states*. Beirut: UNESCO
- Basílio, C. S., Puccini, R. F., Silva, E. M., & Pedromônico, M. R. (2005). Living conditions and receptive vocabulary of children aged two to five years. *Revista de saude publica*, 39, 725-730.
- Bates, E., Andonova, E., D'Amico, S., Jacobsen, T., Kohnert, K., Lu, C., ... & Iyer, G. (2000). Introducing the CRL international picture-naming project (CRL-IPNP). *Center for Research in Language Newsletter*, 12(1), 12-1.
- Bates, E., Benigni, L., Bretherton, L.I., Camioni, L., & Volterra, V. (1979). Cognition and communication from 9-13 months: Correlational findings. In E. Bates (Ed.), *The emergence of symbols: Cognition and communication in infancy*. London: Academic Press.
- Bates, E., Dale, P. and Thal, D. 1995. "Individual differences and their implications for theories of language development". In *Handbook of child language*, Edited by: Fletcher, P. and MacWhinney, B. 96–151. Oxford: Basil Blackwell.
- Berri, H. M., & Al-Hroub, A. (2016). Researching Lebanese Teachers' Knowledge and Perceptions of ADHD. In H. M Berri and A. Al-Hroub, *ADHD in Lebanese Schools* (pp. 21-28). Springer, Cham.

- Berry, J. W. (1990). Acculturation and adaptation: A general framework. In W. H. Holtzman & T. H. Bornemann (Eds.), *Mental health of immigrants and refugees* (pp. 90–102). Austin, TX: Hogg Foundation for Mental Health.
- Berry, J. W. (2002). *Cross-cultural psychology: Research and applications*. Cambridge University Press.
- Bialystok, E. (2007). Acquisition of literacy in bilingual children: A framework for research. *Language Learning*, 57, 45-77.
- Binder, J. R., McKiernan, K. A., Parsons, M. E., Westbury, C. F., Possing, E. T., Kaufman, J. N., & Buchanan, L. (2003). Neural correlates of lexical access during visual word recognition. *Journal of Cognitive Neuroscience*, 15(3), 372-393.
- Bond, M. H., & van de Vijver, F. J. R. (2010). Making scientific sense of cultural differences in psychological outcomes: Unpackaging the Magnum Mysterium. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 75-100). New York, NY: Cambridge University Press.
- Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, 35(1), 158-167.
- Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: Some considerations. *Paidéia (Ribeirão Preto)*, 22(53), 423-432.

- Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, 42(2), 481-487.
- Triandis, H. C., & Brislin, R. W. (1984). Cross-cultural psychology. *American psychologist*, 39(9), 1006.
- Britannica, E. (2011). Encyclopædia Britannica Online. Encyclopædia Britannica, 2011. *Web. Feb, 10.*
- Brownell, H. H., & Caramazza, A. (1978). Categorizing with overlapping categories. *Memory & Cognition*, 6(5), 481-490.
- Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31(2), 141-154.
- Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1), 85-95.
- Carrow-Woolfolk, E. (1985). *Test for Auditory Comprehension of Language-Revised*. Allen, TX: DLM Teaching Resources.
- Casas, R., Calamia, M., & Tranel, D. (2008). A screening test of English naming ability in bilingual Spanish/English speakers. *Journal of Clinical and Experimental Neuropsychology*, 30(8), 956-966.
- Catroppa, C., & Anderson, V. (2004). Recovery and predictors of language skills two years following pediatric traumatic brain injury. *Brain and Language*, 88(1), 68-78.
- Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, 11(41), 63-65.

- Cattell, R. B. (1987). *Intelligence: Its structure, growth and action* (Vol. 35). Amsterdam: North-Holland.
- Catts, H. W., Herrera, S., Nielsen, D. C., & Bridges, M. S. (2015). Early prediction of reading comprehension within the simple view framework. *Reading and Writing, 28* (9), 1407-1425.
- Center for Educational Research and Development. (2018). Retrieved from <https://www.crdp.org/stat-details?id=26000&la=en>
- Champion, T. B., Hyter, Y. D., McCabe, A., & Bland-Stewart, L. M. (2003). "A matter of vocabulary" performances of low-income African American head start children on the Peabody Picture Vocabulary Test—III. *Communication Disorders Quarterly, 24*(3), 121-127.
- Church, A. T. (1987). Personality research in a non-Western culture: The Philippines. *Psychological Bulletin, 102* (2), 272.
- Constable, A., Stackhouse, J., & Wells, B. (1997). Developmental word-finding difficulties and phonological processing: The case of the missing handcuffs. *Applied Psycholinguistics, 18*(4), 507-536.
- Crystal, D. (2004). The Past, Present, and Future of World English. In: A. Gardt and B. Hüppauf (eds) *Globalization and the Future of German*, 27-45. Berlin/New York: Mouton de Gruyter.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222-251.
- Cycowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology, 65*(2), 171-237.

- Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic patients. *Psychological Review*, *104*, 801–838.
- Dollaghan, C. A. (1987). Fast mapping in normal and language-impaired children. *Journal of Speech and Hearing Disorders*, *52*(3), 218-222.
- Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior research methods*, *45*(4), 1246-1258.
- Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808-816.
- Duncan, D., Gibbs, D., Noor, N. S., & Whittaker, H. M. (1988). *Sandwell bilingual screening assessment scales for expressive Panjabi and English*. Windsor: NFER-Nelson.
- Dunn, A. L., & Tree, J. E. F. (2009). A quick, gradient bilingual dominance scale. *Bilingualism: Language and Cognition*, *12*(3), 273-289.
- Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test: PPVT-4A*. Minneapolis, MN: NCS Pearson.
- Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement*, *60*(3), 340-360.
- El Hassan, K., & El Sader, M. (2005). Adapting and validating the BarOn EQ-i: YV in the Lebanese context. *International Journal of Testing*, *5*(3), 301-317.

- El Hassan, K., & Jammal, R. (2005). Validation and development of norms for the Test for Auditory Comprehension of Language-Revised (TACL-R) in Lebanon. *Assessment in Education*, 12(2), 183-202.
- Erkut, S., Alarcón, O., Coll, C. G., Tropp, L. R., & García, H. A. V. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology*, 30(2), 206-218.
- Etard, O., Mellet, E., Papathanassiou, D., Benali, K., Houdé, O., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). Picture naming without Broca's and Wernicke's area. *Neuroreport*, 11(3), 617-622.
- Ferguson, C. A. (1959). Diglossia. *Word*, 15(2), 325-340.
- Fiez, J. A., & Tranel, D. (1997). Standardized stimuli and procedures for investigating the retrieval of lexical and conceptual knowledge for actions. *Memory & Cognition*, 25(4), 543-569.
- Frayha, N. (2009). The negative face of the Lebanese education system. Available online at: <http://www.Lebanonrenaissance.Org/assets/Uploads/0-the-negative-face-of-the-Lebaneseeducation-system-by-Nmer-Frayha-2009.Pdf>.
- Frisbie DA. Essentials of educational measurement. 5th edition. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1991:100-113.
- Games P A. 1984. Data transformations, power, and skew: a rebuttal to Levine and Dunlap. *Psychol. Bull.* 95:345-47
- Gathercole Mueller, V. C. (2013b). *Solutions for the assessment of bilinguals*. Bristol, UK: Multilingual Matters.

- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 304.
- Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.
- Georgas, J. (2003). Family: Variations and Changes Across Cultures. Online Readings in Psychology and Culture, 6(3). <https://doi.org/10.9707/2307-0919.1061>
- Gertner, B. L., Rice, M. L., & Hadley, P. A. (1994). Influence of communicative competence on peer preferences in a preschool classroom. *Journal of Speech, Language, and Hearing Research, 37*(4), 913-923.
- Ghasisin, L., Yadegari, F., Rahgozar, M., Nazari, A., & Rastegarianzade, N. (2015). A new set of 272 pictures for psycholinguistic studies: Persian norms for name agreement, image agreement, conceptual familiarity, visual complexity, and age of acquisition. *Behavior Research Methods, 47* (4), 1148-1158.
- Giddan, J. J., & Milling, L. (1999). Comorbidity of psychiatric and communication disorders in children. *Child and adolescent psychiatric clinics of North America, 8*(1), 19-36.
- Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*(4), 395-427.
- Glaser, W. R. (1992). Picture naming. *Cognition, 42*(1), 61-105.
- Gollan, T. H., Fennema-Notestine, C., Montoya, R. I., & Jernigan, T. L. (2007). The bilingual effect on Boston Naming Test performance. *Journal of the International Neuropsychological Society, 13*(2), 197-208.

- Gollan, T. H., Salmon, D. P., Montoya, R. I., & Galasko, D. R. (2011). Degree of bilingualism predicts age of diagnosis of Alzheimer's disease in low-education but not in highly educated Hispanics. *Neuropsychologia*, *49*(14), 3826-3830.
- Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*(3), 594-615.
- Gottfried, A. W., & Bathurst, K. (1983). Hand preference across time is related to intelligence in young girls, not boys. *Science*, *221*(4615), 1074-1076.
- Gopaul-McNeil, S., & Brice-Baker, J. (1998). *Cross-cultural practice*. New York: Wiley.
- Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, *1*, 131–149.
- Gudykunst, W. B. (1997). Cultural variability in communication: An introduction. *Communication research*, *24*(4), 327-348.
- Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, *9*, 54–65.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In *Adapting educational and psychological tests for cross-cultural assessment* (pp. 15-50). Psychology Press.
- Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education*, *7*(3), 171-186.

- Hammer, C. S., Farkas, G., & Maczuga, S. (2010). The language and literacy development of Head Start children: A study using the Family and Child Experiences Survey database. *Language, Speech, and Hearing Services in Schools, 41*(1), 70-83.
- He J, van de Vijver F (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture 2*. <http://dx.doi.org/10.9707/2307-0919.1111>.
- Helm-Estabrooks, N., Albert, M. L., & Nicholas, M. (2004). *Manual of aphasia and aphasia therapy*. Austin, TX: Pro-ed.
- Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association, 62*(2), 142-147.
- Hoff, E., & Tian, C. (2005). Socioeconomic status and cultural influences on language. *Journal of Communication Disorders, 38*(4), 271-278.
- Holes, C. (2005) _Dialect and national identity: The cultural politics of self representation in Bahrain musalsalaat' in Dresch, P. and Piscatori, J. (eds.) *Monarchies and Nations: Globalization and Identity in the Arab States of the Gulf*, Reading, I.B. Tauris, p. 52-72.
- Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 645–685). New York: Plenum.
- Horn, J. L., & Blankson, A. N. (2012). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 73–98). New

- York, NY: Guilford Press.
- Huff, F. J., Corkin, S., & Growdon, J. H. (1986). Semantic impairment and anomia in Alzheimer's disease. *Brain and Language*, 28(2), 235-249.
- Hutcheon, E. G. (1970). An investigation into stimulus classification under varying instructions. *Unpublished MA thesis, University of Dundee*.
- İlkman, S. (2015). *Reading Acquisition in Primary School-age Children* (Master's thesis, Eastern Mediterranean University (EMU)-DoğuAkdeniz Üniversitesi (DAÜ)).
- International Test Commission. International Test Commission guidelines for translating and adapting tests.2010. https://www.intestcom.org/files/guideline_test_adaptation.pdf Último acceso 06-05-2017.
- International Test Commission. (2018). ITC guidelines in support of the fair and valid assessment of linguistically diverse populations.
- Ives-Deliperi, V. L., & Butler, J. T. (2012). Naming outcomes of anterior temporal lobectomy in epilepsy patients: a systematic review of the literature. *Epilepsy & Behavior*, 24(2), 194-198.
- Jansky, J., & de Hirsch, K. (1972). *Preventing reading failure—Prediction, diagnosis, intervention*. New York: Harper & Row.
- Jusczyk, P. W. (1997). The Discovery of Spoken. *Language, Cambridge: MIT Press*.
- Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test (2nd ed.)*. Philadelphia: Lea & Febiger.
- Kaplan, E., Goodglass, H., Weintraub, S., Segal, O., & van Loon-Vervoorn, A. (2001). Boston naming test: Pro-ed. *Philadelphia: Lea & Febiger*.

- Katz, R. B. (1986). Phonological deficiencies in children with reading disability: Evidence from an object-naming task. *Cognition*, 22(3), 225-257.
- Kaufman, A. S. (2004). *KABC-II: Kaufman Assessment Battery for Children*. AGS Pub.
- Kaufman, J. C. (2009). Creativity, intelligence, and culture: Connections and milieus of creativity. In P. Meusburger, J. Funke, & E. Wunder (Eds.), *Milieus of creativity. An interdisciplinary approach to spatiality of creativity* (Vol. 2, pp. 155—168). Berlin: Springer.
- Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, 47(7), 635-650.
- Khamis, V. (2015). Bullying among school-age children in the greater Beirut area: Risk and protective factors. *Child Abuse & Neglect*, 39, 137-146.
- Khwaileh, T., Body, R., & Herbert, R. (2014). A normative database and determinants of lexical retrieval for 186 Arabic nouns: Effects of psycholinguistic and morpho-syntactic variables on naming latency. *Journal of Psycholinguistic Research*, 43(6), 749-769.
- Kohnert, K. J., Hernandez, A. E., & Bates, E. (1998). Bilingual performance on the Boston Naming Test: preliminary norms in Spanish and English. *Brain and Language*, 65(3), 422-440.
- Kummervold, P. E., Chronaki, C. E., Lausen, B., Prokosch, H. U., Rasmussen, J., Santana, S., & Wangberg, S. C. (2008). eHealth trends in Europe 2005-2007: A population-based survey. *Journal of Medical Internet Research*, 10(4).

- Lecours, A. R. (1975). Myelogenetic correlates of the development of speech and language. *Foundations of Language Development: A multidisciplinary approach, 1*, 121-135.
- Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1-38.
- Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164.
- Little, J., & Ramirez, A. (1976). Ethnicity of subject and test administrator: Their effect on self-esteem. *The Journal of social psychology, 99*(1), 149-150.
- Luyster, R. J., Kadlec, M. B., Carter, A., & Tager-Flusberg, H. (2008). Language assessment and development in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 38*(8), 1426-1438.
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-967.
- Martin, N. A., & Brownell, R. (2011). *Expressive One-Word Picture Vocabulary Test 4*. Novato, CA: Academic Therapy Publications.
- Martin, N., Fink, R. B., Renvall, K., & Laine, M. (2006). Effectiveness of contextual repetition priming treatments for anomia depends on intact access to semantics. *Journal of the International Neuropsychological Society, 12*(6), 853-866.

- Matlock-Hetzel, S. (1997, January). Basic concepts in item and test analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.
- McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136–181). New York, NY: Guilford.
- Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, 53, 258-266.
- Ministry of Education in Lebanon, CRDP: Preliminary Statistics for the Academic Year 2005-2006.
- Montgomery, J. K. (2007). *The bridge of vocabulary: Evidence-based activities for academic success*. Greenville: Pearson Inc.
- Nagy, W. E. (1988). *Teaching vocabulary to improve reading comprehension*. Newark, DE: International Reading Association
- Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labeling. *Journal of Child Language*, 5(1), 1-15.
- Northrup, D. (2005) Globalization and the Great Convergence: Rethinking World History in the Long Term. *Journal of World History* 16(3): 249–267.
- Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17(4), 273-281.
- Oliver, B., Dale, P. S., & Plomin, R. (2004). Verbal and nonverbal predictors of early language problems: An analysis of twins in early childhood back to infancy. *Journal of Child Language*, 31(3), 609-631.

- Padilla, A. (2001). Issues in culturally appropriate assessment. In L.A. Suzuki, J.G. Ponterotto, & P.J. Meller (Eds.), *Handbook of multicultural assessment. Clinical, psychological, and educational applications* (2nd edn., pp. 5–27). San Francisco: Jossey-Bass.
- Paivio, A., Clark, J. M., Digdon, N., & Bons, T. (1989). Referential processing: Reciprocity and correlates of naming and imaging. *Memory & Cognition*, *17*(2), 163-174.
- Panjwani, S. (2012). Developing a naming test for Urdu-English bilinguals: a preliminary study. *Retrieved from University of Texas Digital Repository*.
- Pearson, B. Z. (2008). *Raising bilingual children: A parents' guide*. New York: Random House.
- Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255-1264.
- Piswanger, K. (1975). Cross-cultural comparisons by means of the matrices these of Formann. *German. Unpublished doctoral dissertation, University of Vienna, Vienna*.
- PNUD report, Lebanon toward a citizen state, 2009, p 132.
- Potaka, L., & Cochrane, S. (2004). Developing Bilingual Questionnaires: Experiences from New Zealand in the development of the 2001 Māori language survey. *Journal of Official Statistics*, *20*(2), 289.
- Puente, A. E., Perez-Garcia, M., Vilar-Lopez, R., Hidalgo-Ruzzante, N., & Fasfous, A. F. (2013). Neuropsychological assessment of culturally and educationally dissimilar individuals. In F. Paniagua, & A. M. Yamada (Eds.), *Handbook of*

- multicultural mental health: Assessment and treatment of diverse population* (2nd ed., pp. 225–242). New York: Elsevier.
- Randolph, C., Lansing, A. E., Ivnik, R. J., Cullum, C. M., & Hermann, B. P. (1999). Determinants of confrontation naming performance. *Archives of Clinical Neuropsychology*, *14*(6), 489-496.
- Reynolds, C. R. (2000a). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 249–285). New York, NY: Kluwer Academic/Plenum Press.
- Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. *Handbook of psychology*, 67-93.
- Roberts, P. M., Garcia, L. J., Desrochers, A., & Hernandez, D. (2002). English performance of proficient bilingual adults on the Boston Naming Test. *Aphasiology*, *16*(4-6), 635-645.
- Saeed, A. T., & Fareh, S. (2006). Difficulties encountered by bilingual Arab learners in translating Arabic 'fa' into English. *International Journal of Bilingual Education and Bilingualism*, *9*(1), 19-32.
- Saigh, P. A. (1980). The effects of positive nonverbal examiner comments on the WISC-R performance of Americans in Lebanon. *The Journal of Psychology: Interdisciplinary and Applied*. (104) 165-169.
- Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods*, *6*(2), 169-215.

- Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.). New York, NY: Guilford.
- Serpell, R. (1979). How specific are perceptual skills? A cross-cultural study of pattern reproduction. *British Journal of Psychology*, 70(3), 365-380.
- Serpell, R. (2010). *The significance of schooling: Life-journeys in an African society*. Cambridge University Press.
- Serpell, R., & Deregowski, J. B. (1980). The skill of pictorial perception: an interpretation of cross-cultural evidence. *International Journal of Psychology*, 15(1-4), 145-180.
- Shaaban, K. A. (1997). Bilingual education in Lebanon. *Bilingual Education* (pp. 251-259). Springer Netherlands.
- Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell Labs Technical Journal*, 28(4), 656-715.
- Shawish, H. (2010, June 24). *Campaign to save the Arabic language in Lebanon*. BBC News. Retrieved from <http://www.bbc.com/news/10316914>.
- Simhairi, V. (2010). *An Investigation into the Validity of Adopting Western Psycho-educational Assessments in a Society They are Not Designed for: The Case of Lebanon* (Doctoral dissertation, King's College London).
- Smith, E. E., Balzano, G. J., & Walker, J. (1978). Nominal, perceptual, and semantic codes in picture categorization. *Semantic Factors in Cognition*, 137-168.

- 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 and Memory*, 6(2), 174.
- Snowling, M., Wagtendonk, B., & Stafford, C. (1988). Object-naming deficits in developmental dyslexia. *Journal of Research in Reading*, 11(2), 67-85.
- Soueid, M., Ghanem, S., Hariri, Z., Yamout, N., & Nehme, R. (2014). Analysis of Lebanon's Education Sector. BankMed Market & Economic Research Division. [updated Nov 10; cited 2015 Apr 6]. Available from:
<http://www.bankmed.com.lb/>
- Spinelli, M., Rocha, A. C. D. O., Giacheti, C. M., & Richieri-Costa, A. (1995). Word-finding difficulties, verbal paraphasias, and verbal dyspraxia in ten individuals with fragile x syndrome. *American Journal of Medical Genetics Part A*, 60(1), 39-43.
- Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of education*, 189(1-2), 23-55.
- Super, C. M. (1983). Cultural variation in the meaning and uses of children's intelligence. In J. Deregowski, S. Dziurawiec, & R. Annis (Eds.), *Expiscations in cross-cultural psychology*. Amsterdam: Swets and Zeitlinger.
- Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96, 152-163.
- Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D & Federmeier, K. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, 51(2), 247-250.

- Terrace, H. S. (1985). In the beginning was the" name." *American Psychologist*, 40(9), 1011.
- Tervo, R. C. (2007). Language proficiency, development, and behavioral difficulties in toddlers. *Clinical Pediatrics*, 46(6), 530-539.
- Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, 78(3), 705-722.
- Torres, J. (1991). Equity in education and the language minority student. In *NCBE Forum* (pp. 1-3).
- Tropp, L. R., Erkut, S., Coll, C. G., Alarcón, O., & García, H. A. V. (1999). Psychological acculturation: Development of a new measure for Puerto Ricans on the US mainland. *Educational and Psychological Measurement*, 59(2), 351-367.
- Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 54, 119-135.
- Van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13(1), 29-37.
- Van de Vijver, F.J.R., Leung, K., (1997). *Methods and data analysis for cross-cultural research*. Sage, Newbury Park, CA.
- Van Bakkum, W. J., Houben, J., Sluiter, I., & Versteegh, K. (1997). *The emergence of semantics in four linguistic traditions: Hebrew, Sanskrit, Greek, Arabic* (Vol. 82). John Benjamins Publishing.

- Vogt, P., Douglas, J. M., & Aussems, S. (2015). Early vocabulary development in rural and urban Mozambique. *Child Development Research*, 2015, 15.
doi:10.1155/2015/189195
- Wechsler, D. (1949). Wechsler intelligence scale for children. New York: Psychological Corporation.
- Wechsler, D. (1967). *Wechsler Preschool and Primary Scale of Intelligence-WPPSI*. New York: Psychological Corporation.
- Weder, N. D., Aziz, R., Wilkins, K., & Tampi, R. R. (2007). Frontotemporal dementias: a review. *Annals of General Psychiatry*, 6(1), 15.
- Wiersma, W. & Jurs, S.G. (1990). *Educational measurement and testing* (2nd ed.). Needham Heights, MA: Allyn and Bacon.
- Williams, K. T. (1997). Expressive Vocabulary Test—Second Edition (EVT–2). *Journal of American Academy Child Adolescence Psychiatry*, 42, 864-872.
- Wood, F. B., Hill, D. F., Meyer, M. S., & Flowers, D. L. (2005). Predictive assessment of reading. *Annals of Dyslexia*, 55(2), 193-216.
- Yaghan, M. A. (2008). “Arabizi”: A contemporary style of Arabic slang. *Design Issues*, 24(2), 39-52.
- Zec, R.F., 1993. Neuropsychological functioning in Alzheimer’s disease. In: Parks, R.W., Zec, R.F., Wilson, R.S. (Eds.), *Neuropsychology of Alzheimer’s Disease and Other Dementias*. Oxford University Press, New York, pp. 3–80.
- Zec, R. F., Burkett, N. R., Markwell, S. J., & Larsen, D. L. (2007). A cross-sectional study of the effects of age, education, and gender on the Boston Naming Test. *The Clinical Neuropsychologist*, 21(4), 587-616.

Zeinoun, P., Bawab, S., Atwi, M., Hariz, N., Tavitian, L., Khani, M., Maalouf, F. T.
(2013). Validation of an Arabic multi-informant psychiatric diagnostic interview
for children and adolescents: Development and Well Being Assessment-Arabic
(DAWBA-Arabic). *Comprehensive Psychiatry*, 54(7), 1034-1041.

Appendix A: Rating Booklet Template

Dear Expert Committee Member,

Thank you for being part of this research project! Your time and expertise is contributing significantly to the development of resources that are culturally relevant to our Lebanese students and the children we work with. The research study will acknowledge you as a member of the Expert Committee in the construction of the first Lebanese Picture Naming Vocabulary Test, which will hopefully lead the way to more research in this area.

Task Description: Below is a set of 219 colored pictures. You are kindly asked to:

First: Provide one word that according to you names the picture in Arabic and English, or Arabic and French or all three languages (note: you can write in “chat Arabic” e.g.: teffe7a, kaleb).

Second: On a scale of 1 to 5, rate the level of familiarity of the image to a typical Lebanese child in our community. Familiarity refers to the extent to which we come in contact with or think about the concept/object/animal in our everyday life. *Why? Pictures with low cultural familiarity ratings will be Reviewed from the picture pool.*


Third: On a scale of 1 to 5, rate the level of frequency of the word to a typical Lebanese child in our community between the ages of 3-9 years.

Frequency refers to how often the word occurs in our everyday life whether through conversations, readings or writing. It is important to distinguish between frequency and familiarity. A word can be familiar to our culture but is not encountered frequently such as “tarboush”, “hawiyye”, and “2armed”. *Why? Frequency rating will contribute to the order of the pictures in the test.*

Finally: Based on your experience with children, you are kindly requested to specify the age range during which the word is acquired. *Why? Age of Acquisition rating will contribute to the order of the pictures in the test.*

Deadline: 10 days from today

Note: you can highlight, circle, type or write anywhere in the space, mark the ratings in bold, or add any comment you believe is helpful! I completed the first example for you. You could edit it.

Picture	Name	Image Cultural Familiarity (Rating: 1 to 5)	Word Frequency (Rating: 1 to 5)	Age of Acquisition (years-months)
	Mouse Fara Fa2er Souris	<p>1 = The image is not familiar – a typical Lebanese child will not recognize this image.</p> <p>5 = The Image is very familiar to Lebanese children.</p> <p>Your Rating: 5</p>	<p>1 = Word is uncommon.</p> <p>5 = Word is encountered several times a day in our community.</p> <p>Your Rating: 4</p>	Younger than 3 yrs 3-0 to 3-6 3-7 to 3-11 4-0 to 4-6 4-7 to 4-11 5-0 to 5-6 5-7 to 5-11 6-0 to 6-11 7-0 to 7-11 8-0 to 8-11 9-0 to 9-11 Older than 9 yrs Your Rating: 4-0 to 4-6

Appendix B: Lebanese Psycholinguistic Database for 219 Picture Words

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
1	Ant	animal	Namleh	100.0		4.3	1.2	3.4	0.9	4.6
2	Apple	food	Teffeha	100.0		5.0	0.0	4.9	0.4	2.1
3	Artist	profession	Rassam	83.3	fannan	3.9	1.1	2.6	0.7	5.5
4	Baby	human (or body part) and interaction	Tofol	50.0	walad/bobboo	5.0	0.0	4.9	0.4	2.2
5	Bag	container/receptacle	Kees	100.0		4.6	0.5	4.3	1.4	3.5
6	Baguette	food	Khebez	50.0	rgheef/rgheef franje	4.3	1.2	3.4	1.1	6.1
7	Baloon	toy/game	Balon	100.0		5.0	0.0	4.9	0.4	2.1
8	Banana	food	Mawze	100.0		5.0	0.0	5.0	0.0	2.7
9	Basket	container/receptacle	salleh	100.0		3.9	1.1	3.0	0.8	4.3
10	Beard	human (or body part) and interaction	da2en	66.7	le7iyeh	3.3	1.2	2.6	1.1	6.9
11	Bell	furniture	jaras	100.0		4.1	1.4	4.1	0.8	4.1
12	Belt	clothing (or part of) and accessories	2shat	71.4	zennar	4.4	1.1	3.9	0.9	5.6
13	Bench	furniture	ma23ad	100.0		3.8	1.2	3.0	1.4	5.1
14	Bicycle	vehicle (or part of)	darraje	100.0		4.9	0.4	4.4	0.7	3.7
15	Bin	container/receptacle	zbeleh	83.3	sallet mouhmalet	4.6	0.5	4.6	0.8	3.7
16	Bird	animal	3asfour	100.0		4.9	0.4	4.8	0.7	2.4
17	Bird (Pigeon)	animal	hamama	100.0		3.7	1.5	2.9	1.2	6.9
18	Bone	human (or body part) and interaction	3admeh	100.0		3.6	0.5	2.8	0.9	4.8

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
19	Book	desk/writing material	kteb	100.0		4.9	0.4	4.8	0.5	3.7
20	Boomerang	toy/game	NA	0.0		1.2	0.4	1.2	0.4	9.2
21	Boot	clothing (or part of) and accessories	sobbat	83.3	bot	4.7	0.8	4.3	1.5	3.7
22	Bottle	container/receptacle	2annineh	85.7	zoujehjeh	4.6	0.5	4.3	1.0	4.6
23	Box	container/receptacle	3elbeh	85.7	kartouneh	4.8	0.5	4.0	0.9	4.6
24	Boxer	profession	moulakem	66.7	mousare3	2.4	0.9	1.8	0.9	6.7
25	Brain	human (or body part) and interaction	dmegh	100.0		3.1	1.6	2.1	0.8	8.4
26	Briefcase	container/receptacle	shanta	100.0		3.6	1.0	3.4	1.3	6.5
27	Broccoli	food	broccoli	85.7	arnabeet	2.5	1.2	2.3	1.4	5.9
28	Broom	tools	mekense	100.0		4.0	1.4	3.3	1.0	5.1
29	Butterfly	animal	farashe	100.0		4.9	0.4	4.1	1.0	3.6
30	Button	clothing (or part of) and accessories	zerr	100.0		3.7	1.6	4.2	1.2	3.9
31	Cactus	natural element/plant	sobber	100.0		2.3	1.3	1.6	0.7	7.4
32	Cage	container/receptacle	2afas	100.0		4.3	0.8	2.7	0.8	5.6
33	Camel	animal	jamal	100.0		3.1	1.6	2.5	1.2	6.0
34	Camera	media and communication tools	camera	100.0		3.8	1.2	3.4	0.9	5.4
35	Candle	furniture	chamaa	100.0		4.6	0.5	3.9	0.4	4.3
36	Cap	clothing (or part of) and accessories	bornayta	60.0	2abbou3a/ta2iyye	4.6	0.7	4.0	0.5	4.6
37	Car	vehicle (or part of)	siyyara	100.0		5.0	0.0	5.0	0.0	2.4
38	Carrot	food	jazra	100.0		5.0	0.0	4.5	0.8	3.2

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
39	Chain	tools	janzeer	80.0	selseh	2.9	1.1	2.1	1.0	7.5
40	Chair	furniture	kerse	100.0		5.0	0.0	5.0	0.0	2.9
41	Chair (Armchair)	furniture	kanabeye	66.7	kerse	4.4	0.5	3.3	1.3	6.3
42	Chalkboard	furniture	loh	100.0		4.3	1.4	3.9	1.3	4.2
43	Cheese	food	jebne	100.0		4.3	1.2	4.8	0.5	3.2
44	Chicken	animal	djeje	100.0		5.0	0.0	4.4	0.5	3.8
45	Claws	animal	makhlab	50.0		1.4	1.1	1.2	0.5	8.2
46	Comb	clothing (or part of) and accessories	moshot	100.0		4.1	0.9	3.7	0.8	4.5
47	Compass	desk/writing material	bikar	100.0		2.0	1.4	1.3	0.8	10.2
48	Computer	media and communication tools	7asoub	100.0		3.6	1.5	4.0	1.2	4.5
49	Cone	tools	iqma3 el mourouriah	100.0		2.6	1.5	1.9	1.2	7.3
50	Corn	food	dora	80.4	3arnous	4.4	1.0	3.3	1.0	4.5
51	Cow	animal	ba2ra	100.0		5.0	0.0	4.0	0.8	2.9
52	Crab	animal	salt3oun	100.0		2.8	1.3	2.4	1.1	6.7
53	Cupcake	food	qot3at helo	100.0		4.0	1.1	3.1	0.8	4.3
54	Dart	toy/game	sahem	100.0		2.4	1.3	2.0	0.9	8.5
55	Deer	animal	ghazel	100.0		2.6	1.3	2.1	1.1	5.9
56	Diamond	clothing (or part of) and accessories	almaza	100.0		3.1	1.4	2.0	0.9	6.4
57	Dice	toy/game	zاهر	66.7	7ajar	2.9	1.5	2.3	1.5	6.6
58	Dog	animal	kaleb	100.0		5.0	0.0	4.9	0.4	2.4

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
59	Dolphin	animal	dalfin	100.0		3.6	1.3	2.6	1.2	5.0
60	Dominoes	toy/game	NA	0.0		2.8	1.4	2.1	1.4	6.3
61	Dragon	fiction	tanneen	100.0		2.6	1.3	1.6	0.7	5.8
62	Drums	musical instrument	tabel	100.0		2.6	1.6	1.9	1.1	7.0
63	Duck	animal	batta	100.0		4.4	0.7	3.6	1.1	3.6
64	Envelope	desk/writing material	zaref	80.0	moughallaf	3.0	1.5	2.0	1.3	7.3
65	Face	human (or body part) and interaction	wejj	100.0		4.6	0.5	4.5	0.8	4.1
66	Fan	furniture	marwa7a	100.0		4.5	1.1	4.0	0.9	5.0
67	Finger	human (or body part) and interaction	osba3	100.0		4.5	0.8	3.6	1.3	5.0
68	Fire	natural element/plant	nar	100.0		4.1	0.8	3.6	0.7	3.7
69	Fire extinguisher	tools	taffeye	80.0	metfa2et 7aree2	3.0	1.3	1.6	0.9	8.3
70	Fish	animal	samke	100.0		4.9	0.4	4.5	0.8	2.5
71	Fish Tank	furniture	7od samak	100.0		3.6	1.6	2.3	1.4	6.1
72	Flower	natural element/plant	warde	80.7	zahra	4.9	0.4	4.8	0.5	3.3
73	Foot	human (or body part) and interaction	ejer	100.0		4.9	0.4	3.9	1.4	3.7
74	Football	toy/game	kourat qadam	66.7	tabeh	5.0	0.0	4.5	0.8	4.6
75	Fork	kitchen utensils and appliances	shawke	100.0		5.0	0.0	5.0	0.0	3.1
76	Fruits	food	fweke	100.0		4.9	0.4	4.6	0.7	3.6
77	Gift	container/receptacle	hdiyye	100.0		4.9	0.4	4.1	0.6	3.4
78	Giraffe	animal	zarafeh	100.0		3.5	1.7	2.6	0.9	4.1

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
79	Glass	container/receptacle	kebbeye	100.0		4.4	1.1	4.8	0.5	3.3
80	Glasses	clothing (or part of) and accessories	3waynet	100.0		4.5	0.8	3.8	0.7	4.5
81	Glove	clothing (or part of) and accessories	kfouf	100.0		4.8	0.5	3.9	0.8	3.9
82	Glove (Mitten)	clothing (or part of) and accessories	kfouf	100.0		2.7	1.6	2.6	1.3	5.3
83	Goal	toy/game	shabkeh	100.0		4.1	1.2	3.4	0.5	5.7
84	Goalkeeper	human (or body part) and interaction	golar	100.0		4.1	1.1	3.1	1.1	6.1
85	Goat	animal	me3zeze	83.3	3anze	3.9	1.1	3.0	0.9	5.6
86	Grapes	food	3enab	75.0	tout	3.4	1.6	3.0	1.5	5.8
87	Greenhouse	outdoor places or parts	khaymeh	50.0	al bayt el akhdar	2.0	0.9	1.6	0.7	9.4
88	Guitar	musical instrument	guitar	100.0		4.4	0.7	3.0	1.1	5.2
89	Hair	human (or body part) and interaction	sha3er	100.0		3.9	1.8	4.4	1.4	2.9
90	Hair Brush	tools	fersheye	100.0		4.8	0.5	4.4	0.7	3.6
91	Hamburger	food	hamburger	80.0	burger	5.0	0.0	4.8	0.5	3.2
92	Hammer	tools	shakoush	83.3	matra2a	3.6	1.1	2.8	1.3	6.6
93	Hand	human (or body part) and interaction	eed	100.0		5.0	0.0	4.9	0.4	2.9
94	Hanger	tools	te3li2a	100.0		4.5	0.8	3.3	1.0	6.3
95	Hat	clothing (or part of) and accessories	bornayta	50.0	2abbou3a/ta2iyye	4.5	0.9	3.8	1.2	3.7
96	Heart	shape	aleb	100.0		5.0	0.0	4.3	0.5	3.2

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
97	Helicopter	vehicle (or part of)	merwahiyeh	100.0		3.9	1.4	2.9	1.0	4.3
98	Hippopotamus	animal	7isan el bahr	100.0		2.3	1.2	1.8	0.9	5.9
99	Horse	animal	hsan	100.0		4.5	0.8	3.6	1.1	3.7
100	Hug	human (or body part) and interaction	3abta	66.7	ghamra	4.3	1.1	4.0	0.8	4.2
101	Icecream	food	bouza	100.0		4.9	0.4	4.1	0.6	3.1
102	Island	natural element/plant	jazira	100.0		2.6	1.2	2.0	0.8	6.5
103	Jar	container/receptacle	mortben	100.0		3.8	1.2	3.0	0.9	6.2
104	Key	tools	mefteh	100.0		5.0	0.0	4.6	0.5	3.8
105	Keyboard	media and communication tools	lawhet mafati7	100.0		3.9	1.3	2.9	1.1	6.1
106	Knife	kitchen utensils and appliances	sekkine	100.0		5.0	0.0	4.8	0.5	3.6
107	Lace	clothing (or part of) and accessories	shreet (sobbat)	80.0	7abel	3.8	1.3	4.0	0.5	4.6
108	Lamp	furniture	mosba7 daw	100.0		4.3	0.8	3.5	1.1	5.1
109	Lamp	furniture	lamba	83.3	daw	4.4	1.0	3.9	1.5	5.0
110	Lamp post	outdoor places or parts	daw shere3	33.3	lambet baladiyye/daw	3.1	1.5	2.3	1.1	6.3
111	Leaf	natural element/plant	war2et shajra	100.0		4.3	1.0	3.4	1.0	4.4
112	Lemon	food	hamod	100.0		4.1	0.8	4.0	0.8	4.1
113	Leopard	animal	fahed	100.0		2.1	1.6	1.7	1.0	6.9
114	Lion	animal	assad	100.0		4.3	1.4	3.0	1.1	3.6
115	Lock	tools	2efel	100.0		3.3	1.5	2.6	1.4	7.0
116	Logs	natural element/plant	hatab	100.0		3.5	1.6	2.0	1.1	6.6

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
117	Lumberjack	profession	7attab	100.0		1.5	0.5	1.1	0.4	9.7
118	Maze	toy/game	mata7a	100.0		1.5	0.9	1.4	0.5	8.3
119	Megaphone	media and communication tools	zammour	100.0		2.6	1.1	1.8	1.1	8.5
120	Mermaid	fiction	hourieh	80.0	3arous el bahr	2.1	0.8	1.5	0.8	5.5
121	Microphone	media and communication tools	microphone	100.0		4.0	0.8	2.4	1.2	5.6
122	Microscope	tools	majhar	100.0		2.0	1.0	1.3	0.5	9.1
123	Milk	food	halib	100.0		3.5	1.7	4.0	1.4	3.3
124	Moon	natural element/plant	amar	100.0		4.8	0.5	4.4	0.7	3.4
125	Motorcycle	vehicle (or part of)	darraje nariyye	100.0		4.6	0.7	3.8	0.9	5.1
126	Mouse	animal	fara	100.0		5.0	0.0	4.0	0.0	3.2
127	Mug	kitchen utensils and appliances	fenjen	80.0	kebbeye	4.8	0.5	3.6	0.9	5.1
128	Mushroom	natural element/plant	fotor	100.0		3.6	1.2	2.6	0.7	5.9
129	Necklace	clothing (or part of) and accessories	3a2ed	100.0		3.9	0.8	3.3	1.0	5.5
130	Needle	tools	2ebreh	100.0		3.0	1.4	2.0	1.1	7.0
131	Nose	human (or body part) and interaction	menkhar	100.0		4.5	1.1	4.9	0.4	2.5
132	Notebook	desk/writing material	daftar	100.0		4.5	0.8	4.4	0.5	5.3
133	Orange	food	laymoun	100.0		5.0	0.0	4.5	0.8	3.3
134	Orchestra	profession	fer2a mousi2iyyeh	80.0	orchestra	2.1	1.2	2.0	1.2	8.6

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
135	Pacifier	tools	massasa	100.0		4.9	0.4	3.8	1.0	2.9
136	Paint brush	desk/writing material	fershet	50.0	risheh	4.5	0.8	3.1	1.4	3.8
137	Parachute	vehicle (or part of)	mithalla	100.0		3.1	1.6	2.3	0.9	7.5
138	Parrot	animal	bebbagha2	100.0		3.4	1.3	2.3	0.8	5.7
139	Pear	food	njasa	100.0		4.0	0.8	3.6	0.5	4.4
140	Peg	tools	mal2at (ghassil)	100.0		3.8	1.5	2.9	1.1	6.4
141	Pencil	desk/writing material	2alam	100.0		4.6	0.7	4.9	0.4	3.8
142	Photographer	profession	mousawwer	100.0		3.3	1.5	2.3	1.0	7.4
143	Piano	musical instrument	piano	100.0		4.0	1.1	2.9	1.3	5.0
144	Pineapple	food	ananas	100.0		4.4	0.8	3.6	1.0	4.9
145	Plant	natural element/plant	shatleh	50.0	zarri3a/nabteh	4.9	0.4	3.6	1.1	5.1
146	Pomegranate	food	remmen	100.0		3.1	1.0	2.4	0.7	6.9
147	Pool	toy/game	berkeh	66.7	masba7	4.8	0.5	3.8	0.7	4.1
148	Pot	kitchen utensils and appliances	tanjara	100.0		4.1	1.2	3.4	1.2	6.0
149	Pot (Jug)	kitchen utensils and appliances	jarra	50.0	bree2	1.6	0.9	1.3	0.5	8.2
150	Presto	kitchen utensils and appliances	tanjara (boukar)	100.0		3.1	1.6	3.1	1.1	7.2
151	Puddle	natural element/plant	mayy	50.0	joura	2.1	1.1	3.0	0.8	7.2
152	Puzzle	toy/game	puzzle	100.0		4.5	0.8	3.8	1.0	3.5
153	Rain	natural element/plant	sheta	100.0		4.9	0.4	4.4	0.5	3.6
154	Rectangle	shape	moustateel	100.0		4.2	1.2	3.6	1.0	5.1
155	Rhino	animal	wahid el qarn	100.0		1.9	0.9	1.6	0.8	5.2

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
156	Road	outdoor places or parts	taree2	100.0		4.3	0.8	4.7	0.5	4.5
157	Rooster	animal	deek	100.0		4.5	0.5	3.6	0.9	4.1
158	Rose	natural element/plant	wardeh	100.0		4.8	0.5	4.1	1.1	5.0
159	Runner	human (or body part) and interaction	3adda2	66.7	yarkod	3.9	1.4	3.3	1.5	6.0
160	Safe	kitchen utensils and appliances	brise (kahraba)	100.0		2.8	1.8	2.3	1.2	7.9
161	Salt	food	mele7	100.0		4.3	0.7	4.0	1.1	4.7
162	Saw	tools	menshar	100.0		2.9	1.3	2.0	0.8	8.3
163	Saxophone	musical instrument	NA	0.0		1.8	1.2	1.3	0.5	9.9
164	Scale	tools	mizen	100.0		3.5	0.9	2.0	0.9	7.4
165	Scissors	desk/writing material	m2ass	100.0		5.0	0.0	4.3	0.7	3.6
166	Screen	media and communication tools	telfaz	50.0	7asoub/shesheh	3.6	1.5	4.3	1.2	4.2
167	Screwdriver	tools	mfak bragheh	100.0		2.9	1.6	2.1	1.1	6.4
168	Ship	vehicle (or part of)	safineh	50.0	bekhra/shakhtoura	4.1	1.1	2.9	0.9	4.6
169	Shoe	clothing (or part of) and accessories	sobbat	80.0	7iza2	4.8	0.5	4.6	0.5	3.2
170	Shovel	tools	rafesh	100.0		3.5	1.5	3.0	1.2	5.7
171	Shower	furniture	NA	0.0		3.8	1.4	3.6	1.3	5.1
172	Singer	profession	moughanne	66.7	moutreb/fannan	4.4	0.7	3.4	1.1	6.1
173	Skateboard	toy/game	law7 tazalloj	100.0		3.4	1.2	2.1	1.1	6.3
174	Skeleton	human (or body part) and interaction	haykal 3athmeh	100.0		3.1	1.4	1.9	0.6	7.2

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
175	Snail	animal	bezzay2a	80.0	7alzoun	4.1	1.0	2.9	0.6	4.3
176	Sofa	furniture	kanabeye	85.7	sofa	4.9	0.4	4.5	0.5	4.1
177	Spider	animal	3ankabout	71.4	kertayle	3.8	1.5	2.9	1.3	4.1
178	Spoon	kitchen utensils and appliances	mal32a	100.0		4.9	0.4	4.9	0.4	2.8
179	Square	shape	mourabba3	100.0		4.5	0.5	3.8	1.2	4.1
180	Stadium	outdoor places or parts	mal3ab football	100.0		3.6	1.3	3.1	0.9	6.1
181	Stairs	furniture	daraj	100.0		4.9	0.4	4.5	0.8	3.8
182	Stamp	desk/writing material	tabe3	75.0	khatim	1.3	0.7	1.1	0.4	9.3
183	Stapler	desk/writing material	kebbayse	83.3	debbaseh	4.4	0.7	3.0	0.8	6.5
184	Star	shape	nejmeh	100.0		5.0	0.0	4.1	1.1	3.6
185	Steps	furniture	daraj	71.4	darje	4.3	1.0	4.4	0.7	3.7
186	Stool	furniture	tawleh	100.0		3.0	1.3	2.6	1.4	5.7
187	Strawberry	food	frez	100.0		5.0	0.0	4.3	1.0	3.2
188	Suitcase	clothing (or part of) and accessories	shanta	100.0		3.7	1.4	3.0	0.9	6.0
189	Sun	natural element/plant	shames	100.0		5.0	0.0	4.9	0.4	2.7
190	Sunflower	natural element/plant	douar el shames	66.7	wardeh	3.6	1.6	2.9	1.5	6.0
191	Swimming	human (or body part) and interaction	sabbah	50.0	yasba7/sbe7a	3.9	1.4	3.0	1.1	5.6
192	T-shirt	clothing (or part of) and accessories	blouze	100.0		5.0	0.0	4.9	0.4	3.6
193	Table	furniture	tawleh	83.3	maktab	5.0	0.0	4.9	0.4	3.2

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
194	Tambourine	musical instrument	daff	100.0		3.3	0.9	2.1	1.1	5.9
195	Taxi	vehicle (or part of)	siyyara	100.0		3.6	1.9	3.7	1.5	5.4
196	Teacher	profession	m3allem	50.0	estez	4.1	1.0	4.6	0.5	5.1
197	Teacher (female)	profession	m3allmeh	100.0		4.9	0.4	4.7	0.5	4.1
198	Teapot	kitchen utensils and appliances	bree2	100.0		3.9	1.4	3.0	0.9	5.2
199	Telephone	media and communication tools	telephone	100.0		3.8	1.8	4.6	0.7	3.1
200	Thumb	human (or body part) and interaction	osba3	80.0	ibham	4.9	0.4	3.4	1.2	4.7
201	Tie	clothing (or part of) and accessories	rabtet 3onok	100.0		4.4	0.7	3.3	0.7	6.1
202	Tiger	animal	nemer	80.0		4.1	1.4	3.0	0.9	3.8
203	Tomato	food	banadoura	100.0		5.0	0.0	4.6	0.5	3.2
204	Torch	tools	daw pile	100.0		3.3	1.3	2.9	1.0	6.0
205	Tractor	vehicle (or part of)	NA	0.0		3.9	1.1	3.0	1.1	5.1
206	Train	vehicle (or part of)	train	100.0		3.1	1.6	2.3	1.2	4.7
207	Tray	kitchen utensils and appliances	soniyyeh	100.0		4.3	1.4	3.9	0.4	5.3
208	Tree	natural element/plant	shajra	80.0	arze	4.5	0.5	3.9	0.8	5.0
209	Tree (Round)	natural element/plant	shajra	100.0		5.0	0.0	4.8	0.5	3.0
210	Trophy	toy/game	ka2es	100.0		3.1	1.0	2.3	0.9	7.2
211	Trumpet	musical instrument	bouq	50.0	zammour	2.6	1.5	1.8	0.9	7.4

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Lebanese Psycholinguistic Database for 219 Picture Words

#	Name (English)	Category (superordinate)	Modal Name (Arabic)	Name Agreement %	Alternative words in Arabic	Cultural Familiarity		Word Frequency		Age of Acquisition
						M	SD	M	SD	M
212	Tunnel	outdoor places or parts	nafa2	100.0		3.3	1.0	2.6	1.2	6.2
213	Umbrella	clothing (or part of) and accessories	shamsiyyeh	100.0		4.9	0.4	4.5	0.8	4.1
214	Wasp	animal	na7leh	60.0	dabbour	3.7	0.8	3.1	0.9	4.8
215	Watch	clothing (or part of) and accessories	se3a	100.0		4.8	0.5	4.6	0.5	4.3
216	Wheel	vehicle (or part of)	douleb	100.0		4.9	0.4	3.7	0.5	4.3
217	Wheelbarrow	tools	3arabiyye	100.0		2.8	1.0	1.6	0.9	8.1
218	Xbox	toy/game	NA	0.0		4.1	1.0	3.6	0.7	4.8
219	Zipper	clothing (or part of) and accessories	sa77ab	100.0		3.4	1.8	3.4	1.3	5.6

Note: M = Mean. SD = Standard Deviation. NA = No Lebanese translation was provided.

Appendix C: Decisions to Include, Discard or Examine Items based on Item Parameters after Piloting and Expert's Comments

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Ant	58	Moderate	Very Good	Very Easy	Very Good	Very Easy	Reasonably Good		include
Apple	73	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Artist	43	Very Difficult	Reasonably Good	Easy	Very Good	Very Easy	Marginal	Rater #1: children typically responded as "painting" or "peintre" but did not name the person.	examine
Bag	63	Easy	Very Good	Very Easy	Marginal	Very Easy	Poor		include
Balloon	71	Very Easy	Poor	Very Easy	Very Good	Very Easy	Poor		include
Banana	74	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Basket	58	Easy	Reasonably Good	Easy	Very Good	Very Easy	Poor		include
Bell	56	Moderate	Very Good	Easy	Very Good	Very Easy	Poor		include
Bench	26	Difficult	Very Good	Moderate	Very Good	Difficult	Very Good		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzel, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Bicycle	59	Moderate	Very Good	Very Easy	Poor	Very Easy	Poor		include
Bin	36	Moderate	Poor	Difficult	Poor	Easy	Poor		examine
Bird	69	Very Easy	Reasonably Good	Very Easy	Poor	Very Easy	Marginal		include
Bird (Pigeon)	10	Very Difficult	Poor	Very Difficult	Poor	Very Difficult	Marginal		include
Bone	55	Difficult	Very Good	Very Easy	Very Good	Very Easy	Marginal		include
Book	54	Easy	Very Good	Easy	Marginal	Easy	Poor		include
Boot	27	Difficult	Poor	Difficult	Marginal	Moderate	Poor		include
Bottle	48	Moderate	Reasonably Good	Easy	Very Good	Easy	Marginal		include
Box	46	Moderate	Reasonably Good	Easy	Marginal	Easy	Very Good		include
Brain	42	Very Difficult	Poor	Moderate	Very Good	Very Easy	Very Good	Rater #4: it becomes familiar to students when they learn about it in science class in grade 4.	include
Briefcase	25	Very Difficult	Poor	Difficult	Poor	Moderate	Poor		examine
Broom	45	Difficult	Very Good	Easy	Very Good	Easy	Marginal		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Butterfly	73	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Button	42	Difficult	Very Good	Easy	Very Good	Easy	Very Good	Rater #1: picture is unclear/unfamiliar but the object is very familiar. Rater #2: the size of the button could be misleading. Try reducing it.	examine
Cage	53	Moderate	Very Good	Easy	Very Good	Very Easy	Very Good		include
Camel	37	Very Difficult	Reasonably Good	Moderate	Very Good	Easy	Marginal		include
Camera	49	Difficult	Reasonably Good	Easy	Very Good	Very Easy	Marginal		include
Candle	58	Easy	Very Good	Very Easy	Poor	Very Easy	Reasonably Good		include
Car	73	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Carrot	69	Very Easy	Poor	Very Easy	Very Good	Very Easy	Poor		include
Chair	13	Very Difficult	Poor	Difficult	Very Good	Difficult	Marginal		examine
Chalkboard	64	Easy	Very Good	Very Easy	Marginal	Very Easy	Poor	Rater #8: most schools use whiteboards nowadays.	examine

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Cheese	35	Difficult	Very Good	Moderate	Very Good	Moderate	Reasonably Good	Rater #2: the picture is misleading and it does not look like cheese. It looks like Gruyere.	examine
Chicken	52	Moderate	Poor	Easy	Poor	Easy	Marginal		include
Comb	35	Difficult	Very Good	Moderate	Marginal	Moderate	Poor		include
Computer	41	Difficult	Very Good	Moderate	Very Good	Easy	Very Good	Rater #1: image was a bit misleading. Some said coffee machine or TV.	examine
Corn	60	Easy	Very Good	Easy	Very Good	Very Easy	Marginal		include
Cow	71	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Cupcake	29	Difficult	Very Good	Moderate	Very Good	Difficult	Very Good		include
Diamond	37	Very Difficult	Very Good	Moderate	Very Good	Easy	Very Good		include
Dog	74	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Dolphin	52	Easy	Very Good	Easy	Very Good	Easy	Very Good		include
Duck	63	Very Easy	Reasonably Good	Very Easy	Very Good	Very Easy	Poor		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Envelope	24	Very Difficult	Very Good	Difficult	Very Good	Moderate	Poor		include
Face	54	Moderate	Very Good	Easy	Very Good	Very Easy	Poor	Rater #2: confusing with female figure/woman/girl	examine
Fan	54	Moderate	Very Good	Easy	Marginal	Very Easy	Poor	Rater #1: image is familiar to both English and French speakers but the word is unfamiliar for the French group.	examine
Feather	50	Very Difficult	Very Good	Easy	Marginal	Very Easy	Reasonably Good		include
Finger	58	Easy	Very Good	Easy	Very Good	Very Easy	Reasonably Good		include
Fire	68	Very Easy	Poor	Very Easy	Poor	Very Easy	Marginal		include
Fire Extinguisher	9	Very Difficult	Poor	Very Difficult	Marginal	Very Difficult	Poor		include
Fish	72	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Fish Tank	27	Very Difficult	Poor	Moderate	Very Good	Moderate	Very Good		include
Flower	73	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Foot	70	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Fork	64	Very Easy	Reasonably Good	Easy	Very Good	Very Easy	Poor		include
Fruits	58	Moderate	Very Good	Very Easy	Very Good	Very Easy	Poor		include
Gift	62	Very Easy	Reasonably Good	Very Easy	Poor	Very Easy	Poor		include
Giraffe	65	Very Easy	Reasonably Good	Very Easy	Very Good	Very Easy	Very Good		include
Glass	39	Moderate	Very Good	Moderate	Marginal	Easy	Poor		include
Glasses	70	Very Easy	Poor	Very Easy	Marginal	Very Easy	Poor		include
Glove	42	Moderate	Very Good	Moderate	Very Good	Easy	Very Good		include
Goal	27	Very Difficult	Very Good	Difficult	Poor	Moderate	Very Good		include
Goalkeeper	23	Very Difficult	Poor	Difficult	Very Good	Moderate	Very Good	Rater #3: recognition may depend on the gender of the child.	examine
Goat	42	Difficult	Reasonably Good	Easy	Very Good	Easy	Very Good		include
Guitar	53	Easy	Very Good	Easy	Very Good	Easy	Reasonably Good		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Hair	65	Easy	Very Good	Very Easy	Poor	Very Easy	Poor	Rater #1: typically, children here responded to this as "girl" but could say hair at a younger age if redirected to look at the image. Rater #2: not quite the common hair color or style.	examine
Hair Brush	37	Moderate	Very Good	Easy	Very Good	Moderate	Very Good		include
Hamburger	64	Easy	Very Good	Very Easy	Marginal	Very Easy	Poor	Rater #4: most kids love this kind of fast food (Happy Meals)	include
Hammer	33	Difficult	Very Good	Easy	Very Good	Moderate	Very Good		include
Hand	73	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Hanger	38	Difficult	Very Good	Moderate	Very Good	Easy	Poor		include
Heart	69	Easy	Very Good	Very Easy	Poor	Very Easy	Poor		include
Helicopter	46	Easy	Very Good	Moderate	Very Good	Easy	Very Good		include
Horse	65	Very Easy	Poor	Very Easy	Marginal	Very Easy	Poor	Rater #8: confused for a mule	examine

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Ice Cream	72	Very Easy	Poor	Very Easy	Marginal	Very Easy	Poor		include
Jar	32	Very Difficult	Very Good	Easy	Very Good	Moderate	Very Good		include
Key	68	Easy	Reasonably Good	Very Easy	Poor	Very Easy	Poor		include
Keyboard	15	Very Difficult	Poor	Difficult	Marginal	Difficult	Very Good	Rater #2: with letters and numbers, it might be clearer. Rater #3: picture is not clear. This might be confused with “bricks” or “wall” on a visual level.	include
Knife	70	Very Easy	Reasonably Good	Very Easy	Poor	Very Easy	Poor		include
Lace	13	Very Difficult	Poor	Difficult	Marginal	Difficult	Marginal	Rater #1: object is very familiar but image seems confusing. Rater #2: image is confusing.	examine
Lamp	43	Difficult	Very Good	Easy	Very Good	Easy	Marginal		include
Lamp (Bulb)	53	Easy	Very Good	Moderate	Poor	Very Easy	Poor		include
Leaf	59	Easy	Reasonably Good	Easy	Very Good	Very Easy	Reasonably Good	Rater #1: some children identified as lettuce.	examine

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Lemon	39	Difficult	Very Good	Easy	Very Good	Moderate	Very Good	Rater #2: a full shaped lemon would have been clearer.	examine
Lion	66	Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor		include
Lock	36	Very Difficult	Reasonably Good	Moderate	Very Good	Easy	Marginal		include
Logs	33	Difficult	Very Good	Moderate	Marginal	Moderate	Poor		include
Microphone	49	Difficult	Very Good	Easy	Very Good	Very Easy	Very Good		include
Milk	24	Very Difficult	Reasonably Good	Difficult	Poor	Moderate	Reasonably Good	Rater #1: word is very common but the image was not easily recognized. Rater #3: culturally in Lebanon, milk is more commonly found packed in cartoons or powder milk.	examine
Moon	69	Very Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor		include
Motorcycle	64	Easy	Reasonably Good	Very Easy	Poor	Very Easy	Poor		include
Mouse	61	Easy	Reasonably Good	Easy	Poor	Very Easy	Poor		include
Mug	31	Very Difficult	Reasonably Good	Moderate	Very Good	Moderate	Very Good		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Mushroom	47	Difficult	Very Good	Easy	Very Good	Easy	Very Good		include
Necklace	54	Moderate	Very Good	Very Easy	Poor	Easy	Poor		include
Needle	32	Very Difficult	Poor	Moderate	Marginal	Moderate	Reasonably Good		include
Nose	66	Easy	Very Good	Very Easy	Marginal	Very Easy	Poor		include
Notebook	24	Difficult	Poor	Moderate	Marginal	Difficult	Very Good		include
Orange	69	Very Easy	Reasonably Good	Very Easy	Poor	Very Easy	Poor		include
Pacifier	29	Moderate	Reasonably Good	Moderate	Very Good	Difficult	Marginal		include
Parachute	27	Difficult	Very Good	Difficult	Marginal	Moderate	Very Good		include
Parrot	41	Difficult	Reasonably Good	Moderate	Marginal	Easy	Very Good		include
Pear	55	Easy	Very Good	Easy	Very Good	Easy	Very Good		include
Peg	31	Moderate	Reasonably Good	Difficult	Poor	Moderate	Poor		include
Pencil	69	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Photographer	4	Very Difficult	Poor	Very Difficult	Marginal	Very Difficult	Marginal	Rater #4: which part should the child name?	examine

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Piano	54	Easy	Very Good	Easy	Very Good	Very Easy	Very Good		include
Pineapple	59	Easy	Very Good	Easy	Very Good	Very Easy	Reasonably Good		include
Pomegranate	40	Difficult	Very Good	Moderate	Poor	Easy	Poor		include
Pot	46	Moderate	Very Good	Moderate	Marginal	Easy	Poor	Rater #1: children preferred to name it in Arabic.	include
Presto	30	Difficult	Poor	Difficult	Poor	Moderate	Poor	Rater #3: First, I thought about "tanjara", then when I saw another item for tanjara, I decided it should be called "presto". Rater #7: Highly likely, that child would not recognize the difference between pressure cooker and regular cooking pot, unless it was identified as a goal.	discard
Puzzle	48	Moderate	Very Good	Easy	Very Good	Easy	Very Good		include
Rain	58	Very Easy	Reasonably Good	Very Easy	Marginal	Easy	Poor	Rater #8: rain or raining?	include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Rectangle	35	Difficult	Very Good	Moderate	Very Good	Moderate	Very Good	Rater #1: image is confusing. Some children said square/rectangle/paper/screen.	examine
Road	47	Moderate	Very Good	Easy	Very Good	Easy	Poor		include
Rooster	36	Moderate	Poor	Moderate	Very Good	Moderate	Very Good	Rater #1: confusing rooster/chicken	examine
Rose	48	Difficult	Reasonably Good	Easy	Poor	Very Easy	Marginal	Rater #1: first answer was flower. When I redirected them they said rose. Rater #2: the young students may call it "wardeh"	examine
Salt	60	Easy	Poor	Easy	Very Good	Very Easy	Poor		include
Scale	19	Very Difficult	Poor	Difficult	Poor	Moderate	Reasonably Good		include
Scissors	69	Very Easy	Poor	Very Easy	Marginal	Very Easy	Marginal		include
Shoe	70	Very Easy	Poor	Very Easy	Marginal	Very Easy	Poor		include
Shovel	31	Difficult	Very Good	Moderate	Very Good	Moderate	Very Good		include
Skateboard	35	Difficult	Very Good	Easy	Very Good	Moderate	Very Good		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Skeleton	37	Very Difficult	Poor	Easy	Very Good	Easy	Very Good	Rater #4: some children will only learn this in science class.	include
Snail	52	Easy	Poor	Easy	Very Good	Easy	Very Good		include
Sofa	61	Very Easy	Reasonably Good	Easy	Marginal	Very Easy	Poor		include
Spoon	62	Very Easy	Poor	Easy	Poor	Very Easy	Marginal		include
Square	54	Very Easy	Poor	Easy	Very Good	Easy	Very Good		include
Stadium	26	Very Difficult	Reasonably Good	Difficult	Marginal	Moderate	Poor		include
Stairs	67	Easy	Very Good	Very Easy	Marginal	Very Easy	Poor		include
Stapler	49	Very Difficult	Very Good	Easy	Poor	Very Easy	Poor		include
Star	71	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
Stool	17	Very Difficult	Reasonably Good	Difficult	Marginal	Difficult	Poor		include
Strawberry	70	Very Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor		include
Suitcase	45	Moderate	Poor	Easy	Marginal	Easy	Poor		include
Sun	73	Very Easy	Poor	Very Easy	Poor	Very Easy	Poor		examine
T-Shirt	26	Very Difficult	Very Good	Moderate	Very Good	Difficult	Very Good		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzel, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Table	68	Very Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor		include
Tambourine	16	Very Difficult	Reasonably Good	Difficult	Very Good	Very Difficult	Very Good	Rater #4: most students may not distinguish between drum and tambourine.	examine
Taxi	34	Difficult	Poor	Moderate	Poor	Moderate	Poor	Rater #3: if you mean "taxi", the picture is not clear.	examine
Teacher (Female)	46	Moderate	Reasonably Good	Easy	Very Good	Easy	Poor		include
Teapot	30	Very Difficult	Poor	Moderate	Marginal	Moderate	Reasonably Good		include
Telephone	66	Very Easy	Poor	Very Easy	Poor	Very Easy	Very Good	Rater #3: the picture is outdated. Maybe a handy phone might be better.	examine
Thumb	41	Moderate	Very Good	Moderate	Very Good	Moderate	Very Good		include
Tie	30	Very Difficult	Poor	Moderate	Very Good	Moderate	Very Good		include
Tiger	57	Easy	Very Good	Easy	Very Good	Very Easy	Poor		include
Tomato	68	Very Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor		include
Torch	20	Very Difficult	Poor	Difficult	Very Good	Difficult	Very Good		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Train	56	Moderate	Very Good	Easy	Marginal	Very Easy	Marginal	Rater #2: cultural familiarity depends on the child's socioeconomic background and education. Rater #4: becomes common when they begin to read textbooks and stories.	include
Tray	30	Very Difficult	Poor	Moderate	Very Good	Moderate	Marginal		include
Tree	71	Very Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor		examine
Tree (Round)	71	Very Easy	Reasonably Good	Very Easy	Marginal	Very Easy	Poor	Rater #2: this tree is clearer than the other one.	include
Trophy	46	Very Difficult	Reasonably Good	Easy	Marginal	Easy	Very Good		include
Tunnel	22	Very Difficult	Poor	Moderate	Very Good	Difficult	Very Good		include
Umbrella	67	Easy	Poor	Very Easy	Poor	Very Easy	Poor	Rater #4: context would help.	examine
Watch	69	Very Easy	Poor	Very Easy	Marginal	Very Easy	Poor		include
Wheel	53	Moderate	Very Good	Easy	Marginal	Very Easy	Poor		include

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Item	N	Ages 3-5		Ages 6-7		Ages 8-9		Comments	Decision
		DIF	DISC	DIF	DISC	DIF	DISC		
Zipper	44	Very Difficult	Reasonably Good	Easy	Very Good	Easy	Very Good	Rater #1: unfamiliar as a picture- many kids said jacket.	examine

Note. N = Number of correct responses, DIF= Item Difficulty Index (Item difficulty index is described as *very easy* if it ranges between 0.81-1.00, *easy* = 0.61-0.80, *average/moderately difficult*=0.41-0.60, *difficult*=0.21-0.40 and *very difficult*= 0.00-0.20 (Hetzl, 1997). DISC = Item Discrimination Index range of indices (Item discrimination index is described as *very good* ≥ 0.40 , *reasonably good* = 0.30-0.39, *marginal item* = 0.20-0.29 and *poor* ≤ 0.19 [Frisbie & Ebel, 1991])

Appendix D: Decision to Retain or Review Items

Item	Reason	AoA	Blind Experts' Decision	Final Decision	Expert's Suggestions
Artist	Poor Drawing	-	Include	Retain	
Button**	Poor Drawing	-	Exclude	Revise	Recommend changing the size.
Chalkboard**	Poor Drawing	-	Exclude	Revise	Draw a smart board or white board.
Cheese**	Poor Drawing	-	Exclude	Revise	Picture does not look like cheese from a supermarket.
Computer**	Poor Drawing	-	Exclude	Revise	
Face	Poor Drawing	-	Include	Retain	
Fan	Poor Drawing	-	Include	Retain	
Goalkeeper**	Poor Drawing	-	Exclude	Revise	Add an arrow pointing at the goalkeeper.
Hair	Poor Drawing	-	Include	Retain	
Horse	Poor Drawing	-	Include	Retain	
Keyboard**	Poor Drawing		Exclude	Revise	Add letters and numbers to the keyboard.
Shoelace**	Poor Drawing	-	Exclude	Revise	Change it into a picture of a shoe with an arrow on the lace.
Leaf	Poor Drawing	-	Include	Retain	
Lemon	Poor Drawing	-	Include	Retain	
Milk**	Poor Drawing	-	Exclude	Revise	Add a milk carton.
Photographer**	Poor Drawing	-	Exclude	Revise	Clarify the image.
Presto**	Poor Drawing	-	Exclude	Revise	Delete the picture.
Rectangle	Poor Drawing	-	Include	Retain	

Note. AoA = Age of Acquisition. * Items Reviewed based on Item Parameters and AoA>3. ** Items Reviewed based on two experts' judgement.

Item	Reason	AoA	Blind Experts' Decision	Final Decision	Expert's Suggestions
Rooster	Poor Drawing	-	Include	Retain	
Rose	Poor Drawing	-	Include	Retain	
Taxi**	Poor Drawing	-	Exclude	Revise	Add a taxi logo.
Telephone**	Poor Drawing	-	Exclude	Revise	Change the illustration to a handy phone.
Train	Poor Drawing	-	Include	Retain	
Tree	Poor Drawing	-	Include	Retain	
Zipper**	Poor Drawing	-	Exclude	Revise	Change the picture to a jacket with an arrow pointing at the zipper.
Apple	DISC I <.19	2.1	-	Retain	
Banana	DISC I <.19	2.7	-	Retain	
Bin	DISC I <.19	3.7	-	Retain	
Briefcase*	DISC I <.19	6.5	-	Revise	Delete the picture.
Butterfly	DISC I <.19	3.6	-	Retain	
Car	DISC I <.19	2.4	-	Retain	
Chair	DISC I <.19	2.9	-	Retain	
Cow	DISC I <.19	2.9	-	Retain	
Dog	DISC I <.19	2.4	-	Retain	
Fish	DISC I <.19	2.5	-	Retain	
Flower	DISC I <.19	3.3	-	Retain	
Foot	DISC I <.19	3.7	-	Retain	
Hand	DISC I <.19	2.9	-	Retain	
Pencil	DISC I <.19	3.8	-	Retain	
Sun	DISC I <.19	2.7	-	Retain	

Item	Reason	AoA	Blind Experts' Decision	Final Decision	Expert's Suggestions
Umbrella*	DISC I <.19	4.1	-	Revise	Add rain to the picture and an arrow pointing to an umbrella.

Note: AoA = Age of Acquisiton. * Items Reviewed based on Item Parameters and AoA>3. ** Items revised based on two experts' judgement.

Appendix E: List of the Final Words in the Lebanese Picture Naming Test

#	Target Name in English	Target Name in Lebanese Arabic*	Target Name in French*	AoA in Years
1	Apple	teffe7a	pomme	2.13
2	Balloon	balon	ballon	2.13
3	Car	siyyara	voiture	2.35
4	Dog	kaleb	chien	2.42
5	Bird	3asfour	oiseau ou rossignol	2.44
6	Nose	menkhar	nez	2.48
7	Fish	samke	poisson	2.51
8	Banana	mawze	banane	2.70
9	Sun	shames	soleil	2.70
10	Spoon	mal32a	cuillere	2.83
11	Pacifier	massasa	tetine	2.88
12	Chair	kerse	fauteuil	2.89
13	Cow	ba2ra	vache	2.93
14	Hand	eed	main	2.93
15	Hair	sha3er	cheveux	2.94
16	Tree (Round)	shajra	arbre	2.99
17	Fork	shawke	fourchette	3.05
18	Ice cream	bouza	corner de glace	3.05
19	Table	tawleh	table	3.15
20	Heart	aleb	coeur	3.18
21	Shoe	sobbat	chaussure	3.18
22	Tomato	banadoura	tomate	3.18

Notes. *Modal response in Lebanese Arabic and French. AoA = Age of Acquisition in years

#	Target Name in English	Target Name in Lebanese Arabic*	Target Name in French*	AoA in Years
23	Carrot	jazra	carotte	3.21
24	Mouse	fara	souris	3.21
25	Hamburger	hamburger	hamburger	3.24
26	Strawberry	frez	fraise	3.24
27	Glass	kebbeye	verre	3.29
28	Flower	wardeh	fleur	3.34
29	Orange	laymoun	orange	3.34
30	Gift	hdiyye	cadeau	3.40
31	Moon	amar	lune	3.40
32	Bag	kees	sac	3.46
33	Puzzle	puzzle	puzzle	3.46
34	Duck	batta	canard	3.55
35	Knife	sekkine	couteau	3.56
36	Scissors	m2ass	ciseaux	3.56
37	Butterfly	farashe	papillon	3.57
38	Fruits	fweke	fruits	3.59
39	Lion	assad	lion	3.59
40	Rain	sheta	pluie	3.59
41	T-shirt	blouze	t-shirt	3.59
42	Hair Brush	fersheye	brosse a cheveux	3.63
43	Star	nejmeh	etoile	3.63
44	Bin	zbeleh	corbeille or poubelle	3.65
45	Fire	nar	feu	3.68
46	Bicycle	darraje	bicyclette or velo	3.69
47	Boot	sobbat	bottes	3.71
48	Foot	ejer	pied	3.73
49	Book	kteb	livre	3.74

Notes. *Modal response in Lebanese Arabic and French. AoA = Age of Acquisition in years

#	Target Name in English	Target Name in Lebanese Arabic*	Target Name in French*	AoA in Years
50	Horse	7san	cheval	3.74
51	Chicken	djeje	poule	3.75
52	Key	mefte7	clef	3.75
53	Pencil	2alam	crayon mine	3.75
54	Stairs	daraj	escalier	3.78
55	Tiger	nemer	tigre	3.84
56	Glove	kfouf	gant	3.88
57	Rooster	deek	coq	4.05
58	Giraffe	zarafeh	giraffe	4.06
59	Lemon	hamod	citron	4.06
60	Teacher (female)	m3allmeh	enseignante/prof	4.09
61	Bell	jaras	cloche	4.13
62	Face	wejj	visage	4.13
63	Sofa	kanabeye	fauteuil	4.13
64	Square	mourabba3	carre	4.13
65	Wheel	douleb	pneu	4.25
66	Basket	salleh	panier	4.31
67	Candle	chamaa	bougie	4.31
68	Helicopter	merwa7iyyeh	helicopter	4.31
69	Snail	bezzay2a	escargot	4.32
70	Cupcake	got3at 7elo	muffin	4.33
71	Watch	se3a	montre	4.34
72	Leaf	war2et shajra	feuille ou basilic	4.39
73	Pear	njasa	poire	4.43
74	Comb	mshot	peigne	4.45
75	Corn	dora	mais	4.46
76	Road	taree2	route	4.46

Notes. *Modal response in Lebanese Arabic and French. AoA = Age of Acquisition in years

#	Target Name in English	Target Name in Lebanese Arabic*	Target Name in French*	AoA in Years
77	Glasses	3waynet	lunettes	4.53
78	Ant	namleh	fourmi	4.59
79	Box	3elbeh	boite	4.59
80	Bottle	2annineh	bouteille	4.63
81	Thumb	osba3	pouce	4.68
82	Train	train	train	4.69
83	Salt	mele7	sel	4.71
84	Bone	3admeh	os	4.78
85	Pineapple	ananas	ananas	4.85
86	Lamp	mosba7 daw	lampe	4.95
87	Finger	osba3	index	4.98
88	Dolphin	dalfin	dauphin	4.99
89	Feather	Risheh	Plume	4.99
90	Piano	piano	piano	4.99
91	Tree	shajra	sapin	4.99
92	Fan	marwa7a	ventilateur	4.99
93	Rose	wardeh	rose	5.03
94	Bench	ma23ad	banc	5.05
95	Mug	fenjen	tasse	5.06
96	Rectangle	moustateel	rectangle	5.07
97	Broom	mekense	balai	5.10
98	Lamp	lamba	lampadaire	5.13
99	Motorcycle	darraje nariyye	moto	5.14
100	Guitar	guitar	guitare	5.18
101	Teapot	bree2	theiere ou cafetiere	5.24
102	Notebook	daftar	cahier	5.27
103	Tray	soniyyeh	plateau	5.30

Notes. *Modal response in Lebanese Arabic and French. AoA = Age of Acquisition in years

#	Target Name in English	Target Name in Lebanese Arabic*	Target Name in French*	AoA in Years
104	Camera	camera	appareil photo	5.38
105	Artist	rassam	peintre	5.45
106	Necklace	3a2ed	collier	5.52
107	Goat	me3zeye	chevre	5.56
108	Cage	2afas	cage	5.60
109	Microphone	microphone	micro	5.64
110	Parrot	bebbagha2	perroquet	5.69
111	Shovel	rafesh	pelle	5.70
112	Stool	tawleh	tabouret	5.73
113	Goal	shabkeh	but	5.74
114	Tambourine	daff	tambourin	5.90
115	Mushroom	fotor	champignon	5.91
116	Torch	daw pile	torche/pile	5.95
117	Camel	jamal	chameau	5.96
118	Pot	tanjara	casserole	5.96
119	Suitcase	shanta	valise	6.02
120	Stadium	mal3ab football	stade de foot	6.08
121	Tie	rabtet 3onok	cravate	6.08
122	Fish Tank	7od samak	aquarium	6.13
123	Jar	mortben	bocal	6.17
124	Tunnel	nafa2	tunnel	6.17
125	Skateboard	law7 tazalloj	sakteboard	6.26
126	Hanger	te3li2a	cintre	6.29
127	Peg	mal2at (ghassil)	pince	6.35
128	Diamond	almaza	diamant	6.44
129	Stapler	kebbayse	agrafeuse	6.47
130	Hammer	shakoush	marteau	6.57

Notes. *Modal response in Lebanese Arabic and French. AoA = Age of Acquisition in years

#	Target Name in English	Target Name in Lebanese Arabic*	Target Name in French*	AoA in Years
131	Logs	7atab	bois	6.64
132	Bird (Pigeon)	7amama	pigeon	6.90
133	Pomegranate	remmen	grenade	6.92
134	Needle	2ebreh	aiguille	7.00
135	Lock	2efel	cadenas	7.04
136	Skeleton	haykal 3athmeh	squelette	7.20
137	Trophy	ka2es	trophee	7.20
138	Envelope	zaref	enveloppe	7.32
139	Scale	mizen	balance	7.38
140	Parachute	mithalla	parachute	7.48
141	Fire extinguisher	taffeye	extincteur d'incendie	8.26
142	Brain	dmegh	cerveau	8.44

Notes. *Modal response in Lebanese Arabic and French. AoA = Age of Acquisition in years