*The Plymouth Student Scientist, 2021, **14**, (1), 545-567*

Reliability of targets in a picture naming task

Adele Conn

Project Advisor: Allegra Cattani, School of Psychology (Faculty of Health: Medicine, Dentistry, and Human Sciences), University of Plymouth, Drake Circus, Plymouth, PL4 8AA.

Abstract

The Words in Game (WinG) test is a tool for assessing language development in children, two versions of this test have been developed, one in Italy (Bello et al., 2012) and one in England (Cattani et al., 2019). The present research consists of two studies aiming to determine the reliability of the targets used in the English WinG cards. Study One uses 17 adult participants to rate each set of cards (Italian and English) on how well they represent target constructs, to see if one set is rated as 'better' than the other. Study Two consists of the WinG test being run with 34 child participants to determine if there is a difference in the scores of the two groups of children when one group is tested using the English cards and the other using the Italian cards. Study One finds that the adults rate the English WinG cards as significantly better than the Italian cards in three subtests, it also finds that there are some significant differences in the ratings of individual cards within the test, this suggests that there will be some differences between the scores of the children. However, in Study Two no significant differences are found between the two sets of cards in any subtest, suggesting that changing stimuli does not influence children in the same way as adults. This therefore demonstrates that the targets used in the English WinG cards are reliable when assessing children and that it is appropriate to use this tool for assessing language development in children.

Keywords: Language development, Words in Game, WinG, children, assessing development, visualisation, noun comprehension, noun production, predicate comprehension, predicate production, psychology

Introduction

Language development is a complex process which requires children to learn a great deal, including the sounds, words, and grammar of their language, these develop over time with the first words usually being produced at one year old. However, we know a lot of language learning takes place before the first word is produced (Alcock, 2017; Goldin-Meadow et al., 1976) and those children continue to be able to comprehend words before they can produce them as they are learning language (Capone, 2007). Chomsky proposed that children are born with an innate language learning system, the Language Acquisition Device (LAD) which enables them to acquire language with ease (Chomsky, 1988), this would explain why similar patterns of language development are seen in all children across cultures. Assessing these patterns allows us to determine if a child is developing language normally, at the same rate as the majority of other children their age, or if they have a delay, which can sometimes be an indicator of a wider developmental problem such as autism (Norbury, 2015). Therefore, it is very important to assess language development in young children to determine whether there are any delays or deficits so that interventions can be implemented where appropriate, for example where there is family history of language disorder (Duff et al., 2015). Some delay however may be normal especially if the child is bilingual (Cattani et al., 2012) so it is important to consider the influence of other factors when investigating delays in language development. We know that children learn language before they are able to produce it (Capone, 2007) so we can test comprehension of words as part of assessing their language development, just because they cannot produce the word it does not mean that they do not have a semantic understanding of what it refers to, this means that assessing receptive vocabulary is an important part of assessing overall language development.

The Words in Game (WinG) test was developed in England in 2019 (Cattani et al, 2019) as an adapted version of the Italian Picture Naming Game (PiNG; Bello et al., 2012). WinG is a tool which is used to assess language development in 19 to 36-month-old children, this is the age at which language is being learned and where it is important to be able to identify any delays or deficits so that interventions can be introduced. WinG assesses noun and predicate knowledge, the predicate subtest is more challenging than the noun subtest, as younger children learn nouns first and produce more nouns than any other type of word (Goldin-Meadow et al., 1976) it may be found that younger children struggle with the predicate subtests of the WinG task. Children at the older end of the age range (from 24 months old) should however be able to complete the entire task.

WinG is a direct tool for assessing both language comprehension and production, it was developed as there are very few tools which directly assess language in younger children. The majority of other tools used to assess younger children rely on parental reports to assess language development, such as the Communicative

Development Inventory (CDI; Fenson et al., 1993; Bates et al., 1994), which requires parents to work through a list of words selecting those that their child can understand and those they can say. The CDI is an effective and reliable tool for assessing language development however there is room for error as the parents may over or under-estimate their child's ability, this is why it is important to have direct tools for assessing language development, especially in younger children, and why WinG was developed to be used in England. WinG is a reliable tool for assessing language development and has been shown to have high internal consistency and external validity (Cattani et al., 2019).

Studies have shown that slight changes in visual stimuli can produce different responses. Wu et al. (2015) demonstrated that the type of visual information (context and content) that is presented to students within a test can influence their scores. Furthermore, the colour of the background words are presented on can influence reading comprehension (Zhang et al., 2007) which demonstrates how something such as the background in an image, which seems minor, can affect the words it elicits which of course is important to be aware of when assessing language development. These studies use adult participants so we cannot be sure that the same effects will be seen when the participants are children, studies have shown that children and adults perform differently in perception and action tasks (Duemmler et al., 2008; Schum et al., 2012) which suggests that there are differences in the way that children and adults perceive stimuli and so changing a visual stimulus may affect children differently than it does adults.

The use of pictures is common when assessing language development and naming ability in children (Davidoff & Masterson, 1995) and is an appropriate method for this task so it is important that the reliability of the targets is considered to ensure that children are not being influenced by differing stimuli between different language assessment tools. However, there appears to be little research on how using different pictures to represent the same construct may affect the answers given by children, this means that we have few expectations about how changing images between conditions may influence the reliability of the targets. Research has shown that older children (54 months) may be more able to use additional information provided by the dimension of depth, than younger children (38 months; Kraynak & Raskin, 1971), this demonstrates how the design of stimuli can influence the perception and therefore the responses of children and suggests that perhaps older children are more sensitive to differing stimuli. There are some aspects of stimuli which have been shown to have a negative effect on the accuracy of a child's responses in naming tasks, mainly the 'physical characteristics of the stimulus items' (Walker et al., 2001). Walker et al. (2001) found that children are more able to name 'higher level picture vocabulary items' when they are presented as three dimensional, rather than two-dimensional, line drawings, demonstrating how the stimulus can have an effect on children's performance. Furthermore, studies have shown that

children prefer brighter colours and more brightly coloured stimuli (Boyatzis & Varghese, 1994) which could mean that if the colours of stimuli are changed between conditions the responses given by the children may differ, this would therefore influence the reliability of the targets. This should all be considered when creating a language assessment tool for use with children as it is of great importance that the targets used in these tools are reliable so that tools give accurate insights into a child's ability and interventions can be implemented where appropriate.

As previously discussed, WinG is a reliable tool (Cattani et al., 2019), however, as the WinG test was adapted from the Italian version of the test (Bello et al., 2012) there are two sets of cards each using different images to represent the same constructs. These visual differences in the stimuli may lead to different, or more accurate, responses being produced by children being tested using one of the sets of cards compared to children being tested on the other set of cards. So, it is important to know that the targets shown on the English WinG cards produce similar responses to the Italian WinG cards and so are reliable for assessing language development in children.

The reliability of the targets in a language assessment tool is very important because it means that the tool has interrelatedness between items and so is measuring language development throughout the whole assessment. This also means that the targets in the assessment have consistency both within the tool as well as when compared to other language assessment tools, so it is only appropriate to use it to assess language development if the targets are reliable.

The present study

In Study One, a pre-test with adult participants will be conducted to determine whether adults think that one set of cards (Italian or English) better represented the constructs than the other set, this will help us determine if one set of cards more clearly represents the constructs, if this was the case then we would expect the children to perform better (produce more correct and non-target but semantically related answers) when using this set of cards. Adults are good participants for this as they are already experts on the words and constructs used in the WinG test and so can give an unbiased idea of whether the stimuli they are presented with match their expectations of how they would imagine the construct and whether the pictures used in WinG clearly represent those constructs.

In Study Two we will be running the Words in Game (WinG) test, with children, as described in the manual (Cattani et al., 2019), the children will either be tested using the Italian cards or the English cards, so that we can investigate whether the two sets of cards produce similar scores or if one elicits significantly more correct responses than the other due to the differences in the images on the cards. We will also be testing children from two age groups, 24 months, and 30 months, in order to see if there are any differences between the cards for different age groups.

Our first hypothesis is that in Study One there will be differences in the ratings of the two sets of WinG cards (English and Italian), we expect that the two sets of cards will be rated differently due to the differences in the images on the cards. We expect this because studies with adult participants have shown that even subtle differences in stimuli can affect the answers that they elicit (Wu et al., 2015; Zhang et al., 2007).

The second hypothesis is dependent on the outcome of Study One. We expect that if one set of cards is rated as significantly better than the other set, the children tested using the 'better' set of cards will produce more correct and semantically related answers than the children being tested using the other set of cards. We expect this because the adults have much more knowledge about the target items than the children do due to their life experience, so if they rate one set of cards as significantly better at representing that target in Study One it should be easier for the children to understand what those cards are representing and so produce more correct answers. Furthermore, if one set of cards is rated as 'better' in Study One we expect that children tested with that set of cards will give more semantically related answers than children tested using the other set of cards. This is because if the adults think the cards represent the construct better, the image should be a clearer representation of that construct. Therefore, children tested using the 'better' set of cards may be more able to access their semantic knowledge of the construct due to the clarity of the image and so will be more able to answer the production question with a semantically related answer, if they do not know the target word, than the children tested using the other set of cards.

If we do not find differences between the two sets of cards in either study it will demonstrate the reliability of the targets used in the English WinG cards because it will show that despite the changes in the images on the cards they still elicit the same responses as the Italian WinG cards.

Method

Materials

The Words in Game (WinG) test consists of 132 coloured picture cards which depict various constructs, designed to test children's knowledge of different types of words, nouns and predicates. Both the Italian and English sets of the WinG cards were used for every participant in study one, and answers were recorded on a spreadsheet. Due to the circumstances surrounding the Covid-19 pandemic, after some data had been collected using the actual cards it was necessary to take photos of them so that images could be sent to participants. The pictures of the cards were transferred to word documents where the two pictures representing the same construct were presented side by side with the construct they were representing written in between the images.

The Words in Game (WinG) cards, both the Italian and English versions, were also used in study two, children were tested on either the Italian cards (N=18) or the English cards (N=16). The WinG cards are presented to the children in triplets consisting of a comprehension card, a distractor card, and a production card. In the first half of the task noun comprehension and production is tested, this assesses basic vocabulary that represents common objects, in the second half of the task the children are tested on comprehension and production of predicates, which assesses more complex vocabulary development including knowledge of verbs, adjectives, and adverbs. A copy of the short form Macarthur-Bates CDI was filled out by the parents of the children taking part in this study. The Words in Game record form was used during the study by the coder and each session was recorded using a camera on a tripod in the corner of the booth.

Study One

Participants

17 adults, aged 18 to 55 (10 males, 7 females) were recruited to take part in the adult studies, some were recruited through the University of Plymouth School of Psychology participation pool where students of psychology can sign up to take part in studies, other participants were friends and family of the researchers who volunteered to take part.

Procedure

For some, the study took place in a quiet room with the researcher and the participant sat at a table facing each other, other participants were sent the documents containing the images and the study took place over a video call, due to the circumstances surrounding Covid-19. The researcher explained the study and what the participant would be required to do and gained their informed consent to take part in the study. Starting with the noun cards before moving onto the predicate cards, the researcher presented the cards from both sets of the WinG task two at a time so that cards representing the same constructs were presented together and asked the participant to rate them on a scale of 1-4 (so that there was no option for a neutral rating) whilst reminding them not to compare the cards but simply to rate them on how well they represented the construct. The study continued in this way until all the cards (from both sets) had been presented and rated. Once the study was complete the participant was thanked (students of the school of psychology were given a participation point for their time) and given the opportunity to ask any questions they had, this study involved no deception but a debrief was given to explain what the researchers were doing and expecting to find.

Results: Study One

For each card, the mean of all the ratings was calculated and analysed before running any further statistical tests. A paired samples t-test was then conducted to

determine whether there were any actual and significant differences between ratings of individual cards as well as each subsection overall.

Noun Comprehension

For the noun comprehension subtest there appeared to be a difference between the following cards: Mountain (English card *M*=3.94, Italian card *M*=2.94), Backyard (English card *M*= 3.41, Italian card *M*=2.47), and Bib (English card *M*=3.88, Italian card *M*=3.18). However, the overall means for the two sets of cards in this subsection appeared to be very similar (English cards *M*=3.88, Italian cards *M*=3.63 (see Table 1).

Table 1: Shows the mean and standard deviations (SD) of individual cards and the noun comprehension subtest overall.

	English cards		Italian cards	
	Mean	SD	Mean	SD
Overall	3.88	.14	3.63	.34
Mountain	3.94	.24	2.94	.83
Motorbike	4	0	3.71	.47
Iron	3.94	.24	3.59	.62
Sofa	4	0	3.65	.49
Clouds	3.88	.49	3.47	.72
Backyard	3.41	.71	2.47	.94
Bib	3.88	.33	3.18	.64

A paired samples t-test showed a statistically significant difference between the ratings of the two sets of cards with the English cards (*M*=3.88, *SD*=0.14) being rated as better than the Italian cards (*M*=3.63, *SD*=0.34), $t(16) = 3.647$, $p = .002$. Furthermore, some differences between the images were found to be significant these were for the following cards: Mountain, the English card (*M*=3.88, *SD*= .14) was rated as better than the Italian card (*M*=3.63, *SD*= .35), $t(16)=4.76$, $p<0.001$. Motorbike, the English card (*M*=4, *SD*= 0) was rated as better than the Italian card (*M*=3.71, *SD*= .47), $t(16)=2.58$, $p= .02$. Iron, the English card (*M*=3.94, *SD*= .24) was rated better than the Italian card (*M*=3.59, *SD*= .62), $t(16)=2.95$, $p= .009$. Sofa, The English card (*M*=4, *SD*= 0) was rated better than the Italian card (*M*=3.65, *SD*= .49), $t(16)=2.95$, $p= .009$. Clouds, the English card (*M*=3.88, *SD*= .49) was rated as better than the Italian card (*M*=3.47, *SD*= .72), $t(16)=2.75$, $p= .14$. Backyard, the English

card (*M*=3.41, *SD*= .71) was rated as better than the Italian card (*M*=2.47, *SD*= .94), *t*(16)=3.11, *p*= .007. Finally, Bib, the English card (*M*=3.88, *SD*= .33) was rated as better than the Italian card (*M*=3.18, *SD*= .64), *t*(16)=4.24, *p*= .001.

Noun Production

For the noun production subtest, the cards which looked to have been rated as being different were Book (English card *M*=3.47, Italian card *M*=2.88), and Radiator (English card *M*=3.94, Italian card *M*=3.41). The overall mean ratings for the two sets of cards did not look to be very different (English cards *M*=3.87, Italian cards *M*=3.65), see Table 2.

A paired samples t-test showed that the two sets of cards were rated as being significantly different, the English cards (*M*=3.87, *SD*= .20) were rated overall as better than the Italian cards (*M*=3.65, *SD*= .36), *t*(16)=3.47, *p*= .003, demonstrating that the adults rated the English cards as representing the constructs significantly better than the Italian cards. There were also significant differences between the following cards: Hen, the English card (*M*=4, *SD*= 0) was rated as significantly better than the Italian card (*M*=3.76, *SD*= .44), *t*(16)=2.22, *p*= .04. Socks, the English card (*M*=4, *SD*= 0) was rated as better than the Italian card (*M*=3.65, *SD*= .61), *t*(16)=2.4, *p*= .029. Roof, the English card (*M*=3.65, *SD*= .79) was rated as significantly better than the Italian card (*M*=3.24, *SD*= .75), *t*(16)=2.38, *p*= .03. Glass, the English card (*M*=4, *SD*= 0) was rated as better than the Italian card (*M*=3.65, *SD*= .61), *t*(16)=2.4, *p*= .029. Finally, Radiator, the English card (*M*=3.94, *SD*= .24) was rated as better than the Italian card (*M*=3.41, *SD*= .8), *t*(16)=2.73, *p*= .015.

Table 2: Show the means and standard deviations (SD) of individual cards and the noun production subtest overall.

	English cards		Italian cards	
	Mean	SD	Mean	SD
Overall	3.87	.20	3.65	.36
Hen	4	0	3.76	.44
Socks	4	0	3.65	.61
Roof	3.65	.79	3.24	.75
Glass	4	0	3.65	.61
Radiator	3.94	.24	3.41	.8

Predicate Comprehension

For the predicate comprehension (Table 3) subtest Swinging (English card *M*=3.82, Italian card *M*=3), and Full (English card *M*=2.35, Italian card *M*=2.82) appeared to have a large difference in the average ratings. Additionally, some cards representing adjectives have a much lower average, for both sets of cards, than other cards in the subtest, these are Big (English card *M*=1.47, Italian card *M*=1.41), and Outside (English card *M*=1.24, Italian card *M*=1.59) which suggests that the children may struggle more with these cards regardless of the set used. The overall average for each set of cards did not appear to be very different from one another (English cards *M*=3.17, Italian cards *M*=3.08).

A paired samples t-test showed no significant difference between the English set of cards (*M*=3.17, *SD*= .42) and the Italian set of cards (*M*=3.08, *SD*= .38), *t*(16)=1.38, *p*= .187, see Table 3. However, there were some significant differences between the following cards: Swinging: the English card (*M*=3.82, *SD*= .39) was rated as better than the Italian card (*M*=3, *SD*= .94), *t*(16)=3.57, *p*= .003. Hugging, the English card (*M*=3.47, *SD*= .712) was rated as worse than the Italian card (*M*=3.88, *SD*= .33), *t*(16)=-2.75, *p*= .014. Behind, the English card (*M*=2.94, *SD*= .97) was rated as better than the Italian card (*M*=2.59, *SD*= .87), *t*(16)=2.95, *p*= .009. Building, the English card (*M*=3.41, *SD*= .87) was rated as better than the Italian card (*M*=3.18, *SD*= .95), *t*(16)=2.22, *p*= .041, see Table 3.

Table 3: Show the means and standard deviations (SD) of individual cards and the predicate comprehension subtest overall.

	English cards		Italian cards	
	Mean	SD	Mean	SD
Overall	3.17	.42	3.08	.38
Swinging	3.82	.39	3	.94
Hugging	3.47	.72	3.88	.33
Behind	2.94	.97	2.59	.87
Building	3.41	.87	3.18	.95

Predicate Production

Finally, for the predicate production subtest there was one card which had a lower average for both the English card (*M*=1.76) and the Italian card (*M*=1.88) this was

Small, and so this may also be a card that children struggle with regardless of the card set used.

There appeared to be differences in the mean ratings of the following cards (Table 4): Spinning (English card *M*=3.71, Italian card *M*=2.59), Falling (English card *M*=3.76, Italian card *M*=2.94), and Opening (English card *M*=3.71, Italian card *M*=3.29). Again, the overall averages for each set of cards in this subtest appeared very similar (English cards *M*=3.29, Italian cards *M*=3.11).

A paired samples t-test (also see Table 4) showed that overall there was a significant difference between the English set of cards (*M*=3.29, *SD*= .42) and the Italian set of cards (*M*=3.11, *SD*= .43), *t*(16)=2.72, *p*= .015, which shows that the adults rated the English set of cards for this subtest as better at representing the construct than the Italian set of cards, as detailed in Table 4. There are also significant differences between the following cards: Spinning, the English card (*M*=3.71, *SD*= .59) was rated as better than the Italian card (*M*=2.59, *SD*= .8), *t*(16)=4.97, *p*< .001. Clean, the English card (*M*=2.53, *SD*=1.01) was rated as better than the Italian card (*M*=2.29, *SD*= .99), *t*(16)=2.22, *p*= .041). Falling, the English card (*M*=3.76, *SD*= .44) was rated as better than the Italian card (*M*=2.94, *SD*= .9), *t*(16)=4.67, *p*< .001. Smiling, the English card (*M*=3.94, *SD*= .24) was rated as better than the Italian card (*M*=3.71, *SD*= .47), *t*(16)=2.22, *p*=0.041. Lastly, Opening, the English card (*M*=3.71, *SD*= .47) was rated as better than the Italian card (*M*=3.29, *SD*= .77), *t*(16)=2.14, *p*= .049.

Table 4: Show the means and standard deviations (SD) of individual cards and the predicate production subtest overall.

	English cards		Italian cards	
	Mean	SD	Mean	SD
Overall	3.29	.42	3.11	.43
Spinning	3.71	.59	2.59	.8
Clean	2.53	1.01	2.29	.99
Falling	3.76	.44	2.94	.9
Smiling	3.94	.24	3.71	.47
Opening	3.71	.47	3.29	.77

This suggests that we may find some differences between the scores from children tested using the Italian cards and those tested using the English cards in Study Two. Furthermore, differences in the ratings of individual cards suggests that we may see differences between these cards in the children. Overall, these findings support our first hypothesis and therefore suggest we should test our second hypothesis.

Method: Study Two

Participants

The participants in Study Two were 19 children aged 30 months (7 males, 12 females) and 15 children aged 24 months (9 males, 6 females) who were all recruited by the Babylab at the University of Plymouth. All children had normal or corrected vision, no hearing deficits (as far as the parents were aware) and were not born more than six weeks prematurely. The educational levels of the parents of the children in this study ranged from the equivalent of A-levels to postgraduate qualifications.

Procedure

Firstly, the children were recruited by the Babylab at the University of Plymouth and parents were invited to bring their child into the Babylab on a day to suit them when the child was 24 or 30 months old (within the month after they turned 24 or 30 months old and before they turned 25 or 31 months old). The card set to be used was selected and placed into the booth before the child arrived at the Babylab.

Upon entering the Babylab the children were invited to play as this gave them a chance to become familiar with the experimenter and the environment, meanwhile parents were given an information sheet explaining what their child would be doing, a consent form, and were asked to fill out the Macarthur-Bates CDI, if it was their first visit to the Babylab they were also given a demographics form to complete. Then the child and their parent, the experimenter, and a coder moved into a separate booth in the Babylab for the study to begin. The child and the experimenter sat at a small table in the centre of the room with the camera pointing towards the child so that the session could be re-watched to code any words the coder missed or was unsure of, the parent was sat behind the child however did have the option to move closer to the child if that helped the child to feel more comfortable in the environment, the coder sat to the left of the experimenter so they could see the child and the cards he/she was pointing to.

The experiment began with two practice trials where the first triplet of WinG cards were presented to the child (placed on the table in a line, but random order) and the experimenter was able to explain what they wanted the child to do without having to keep to the phrases suggested in the WinG manual (Cattani et al., 2019) as in the real trials, this is so that the experimenter could ensure that the child understood the game and felt comfortable doing the study. In the first practice trial pictures of a cat

(comprehension), a dog (production), and a television (distractor) were presented and the child was first asked the noun comprehension question – to point to the cat, only one attempt is allowed at this unless the child spontaneously changes his/her mind in which case the second answer is accepted, the experimenter then takes the cat and the television cards away and asks the child what the picture on the remaining card is of, if the child does not respond correctly a second chance is given and the best answer is scored. The coder for each WinG session uses the record sheet to write next to the word which card the child pointed to in the comprehension part and the answer (or answers) they gave for the production questions. The game continues in this way until all 22 triplets in the noun sub-section have been completed before moving onto the predicate cards, with a short break between the two sub-tests if the parent or experimenter deem it to be necessary.

For the predicate cards, the same procedure is followed starting with two practice trials, the only thing that differs are the phrases used by the experimenter as suggested in the WinG manual. The study usually takes around 20 to 30 minutes depending on the child, once it is complete the child is given a balloon, a certificate and a small gift to thank them for their participation, the parent also has the opportunity to ask any questions they may have before they leave.

Results: Study Two

Data from children who did not complete the test up to the seventeenth triplet in a subtest was excluded from the analysis on that particular subtest. So, in this analysis we have analysed the data from six 24-month-olds and 18 30-month-olds on the noun comprehension and production subtests (12 with the English cards, 12 with the Italian cards), and two 24-month-olds and 16 30-month-olds on the predicate comprehension and production subtests (10 with the English cards, 8 with the Italian cards).

Demographics

Firstly, the CDI ‘understands’ and ‘says’ scores from the children in both groups were compared to ensure that all children were of similar abilities and that there were no significant differences between the two groups which could influence the scores. A *t*-test confirmed that there were no significant differences between the two groups in the scores for the CDI ‘understands’ (Italian card group M=53.08, SD= 26.47, English card group M=58.33, SD= 21.36), $t(22)$= -.535, p= .598.

Furthermore, there were no significant differences between the scores for the two groups for the CDI ‘says’ (Italian card group M=40.5, SD=24.76, English card group M=44.75, SD= 23.96), $t(22)$= -.427, p= .673. Demonstrating that the two groups are of similar abilities and so any differences we find will be due to the differences between the two sets of WinG cards rather than differing abilities between the groups.

An independent samples t-test showed that there were also no significant differences between genders in any of the sub-tests, as indicated in Table 5.

Table 5. Show the means and standard deviations (SD) of scores on each subtest for Males and Females.

	Males		Females	
	Mean	SD	Mean	SD
Noun Comprehension	15.36	4.37	16.92	2.99
Noun Production	8.91	2.88	11.15	3.93
Predicate Comprehension	14.25	2.55	16.2	1.81
Predicate Production	7.25	3.15	8.4	1.96

A Spearman's correlation was conducted to assess the relationship between the CDI 'understands' score and the score on each subsection of the WinG test. There was no significant correlation between the CDI 'understands' score and the noun comprehension subtest, $r_s(22) = .396$, $p = .055$. There was a significant correlation between the CDI 'understands' score and the overall score on the noun production subtest, $r_s(22) = .500$, $p = .013$. There was no significant correlation between the CDI 'understands' score and the predicate comprehension subtest, $r_s(16) = .027$, $p = .915$, nor was there a significant correlation between the CDI 'understands' score and the predicate production subtest, $r_s(16) = .002$, $p = .995$.

A Spearman's correlation was also conducted to assess the relationship between the CDI 'says' score and the overall score in each subtest. There was a significant correlation between the CDI 'says' score and the noun comprehension subtest, $r_s(22) = .491$, $p = .015$. There was also a significant correlation between the CDI 'says' score and the noun production subtest, $r_s(22) = .771$, $p<.001$. There was no significant correlation between the CDI 'says' score and the predicate comprehension subtest, $r_s(16) = .238$, $p = .341$, nor was there a significant correlation between the CDI 'says' score and the predicate production subtest, $r_s(16) = .260$, $p = .298$.

Comparison of English and Italian cards

As significant differences were found in Study One, a Mann-Whitney-U analysis was conducted on the data from Study Two to determine whether there were any significant differences between the two groups. A Mann-Whitney-U test was used as opposed to a parametric test for these comparisons as it was not appropriate to use a parametric test due to the data violating assumptions especially for the noun sub-

tests. There was only one violation of the assumptions in the predicate sections however, a non-parametric test was conducted for consistency across the data analysis.

Noun sub-tests

As *Figure 1* shows there did not appear to be differences between the mean number of correct responses elicited by each set of cards in the noun subtests. A Mann-Whitney-U test was run to determine if there were any statistically significant differences in the number of correct responses elicited by the English WinG cards and the Italian WinG cards in the noun comprehension sub-test. Distributions of the scores for the English cards and the Italian cards were not similar, as assessed by visual inspection. The number of correct responses for the English cards (mean rank = 15.04) and the Italian cards (mean rank = 9.96) were not statistically significantly different, $U = 41.5$, $z = 1.779$, $p = .078$, using an exact sampling distribution for U (Dineen & Blakesley, 1973).

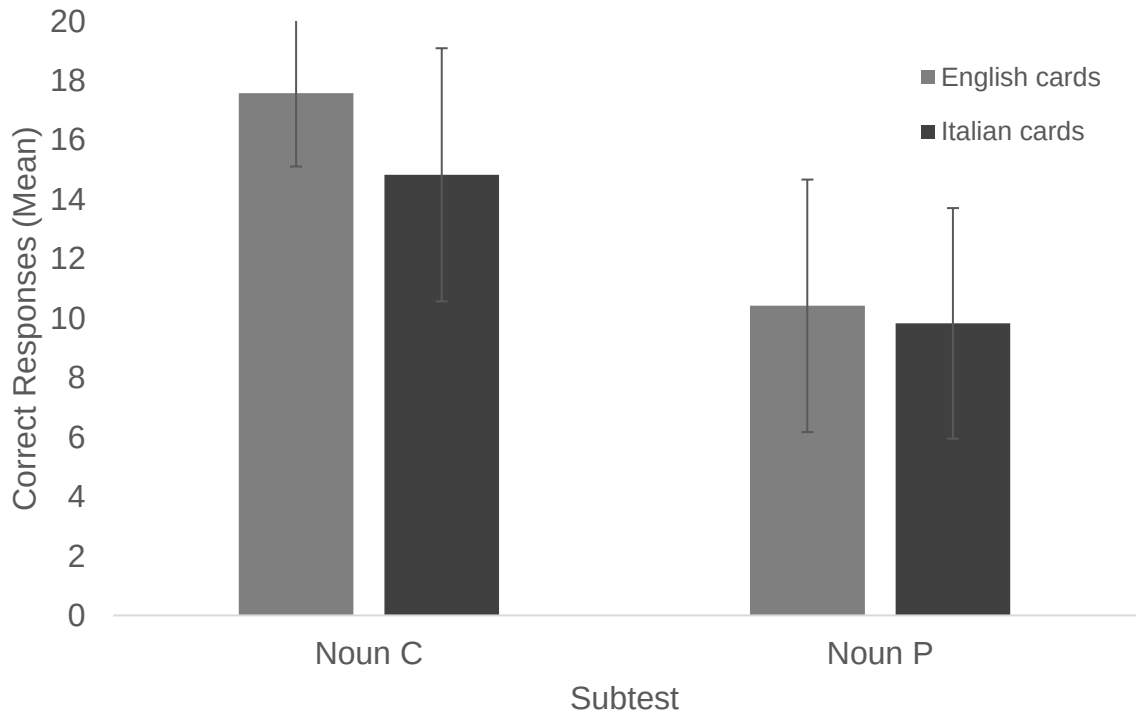

Figure 1: Graph showing the mean of correct responses elicited by each set of WinG cards for the noun comprehension (Noun C) and production (Noun P) subtests.

A Mann-Whitney-U test was also run to determine if there were differences in the number of correct responses elicited by the English WinG cards and the Italian WinG cards in the noun production sub-test. Distributions of the scores for the English cards and the Italian cards were not similar, as assessed by visual inspection. The

number of correct responses for the English cards (mean rank = 13.08) and the Italian cards (mean rank = 11.92) were not statistically significantly different, *U* = 65, *z* = .406, *p* = .713, using an exact sampling distribution for *U* (Dineen & Blakesley, 1973).

Predicate sub-tests

As *Figure 2* shows there was little difference in the means for each set of cards for the predicate subtests. A Mann-Whitney-U test was run to determine if there were any statistically significant differences in the number of correct responses elicited by the English WinG cards and the Italian WinG cards in the predicate comprehension sub-test. Distributions of the scores for the English cards and the Italian cards were not similar, as assessed by visual inspection.

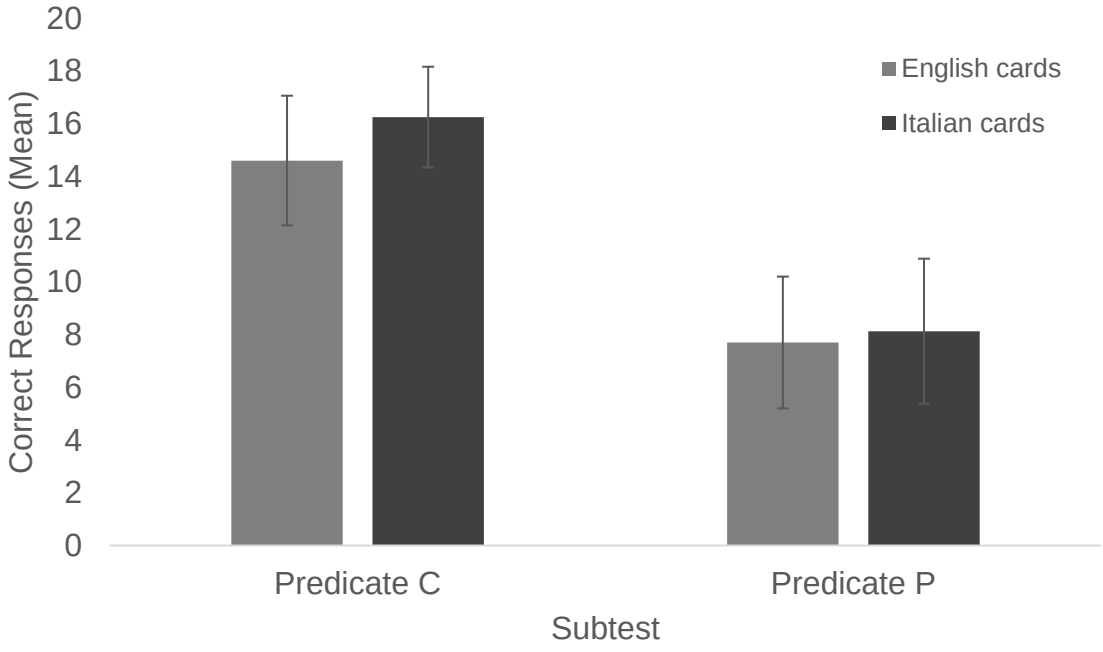

Figure 2: Graph showing the mean of correct responses elicited by each set of WinG cards for the predicate comprehension (Predicate C) and production (Predicate P) subtests

The number of correct responses for the English cards (mean rank =7.95) and the Italian cards (mean rank = 11.44) were not statistically significantly different, *U* = 24.5, *z* = -1.398, *p* = .173, using an exact sampling distribution for *U* (Dineen & Blakesley, 1973).

A Mann-Whitney-U test was also run to determine if there were significant differences in the number of correct responses elicited by the English WinG cards and the Italian WinG cards in the predicate production sub-test. Distributions of the scores for the English cards and the Italian cards were not similar, as assessed by visual inspection. The number of correct responses for the English cards (mean rank = 9.20) and the Italian cards (mean rank = 9.88) were not statistically significantly

different, $U = 37$, $z = -.269$, $p = .829$, using an exact sampling distribution for U (Dineen & Blakesley, 1973).

Non-target but semantically related errors

As *Figure 3* shows, there did not appear to be differences between the number of semantically related responses elicited by each set of cards in both production subtests. A Mann-Whitney-U test was run to determine if there were any statistically significant differences in the number of non-target but semantically related errors elicited by the English WinG cards and the Italian WinG cards in the noun production sub-test.

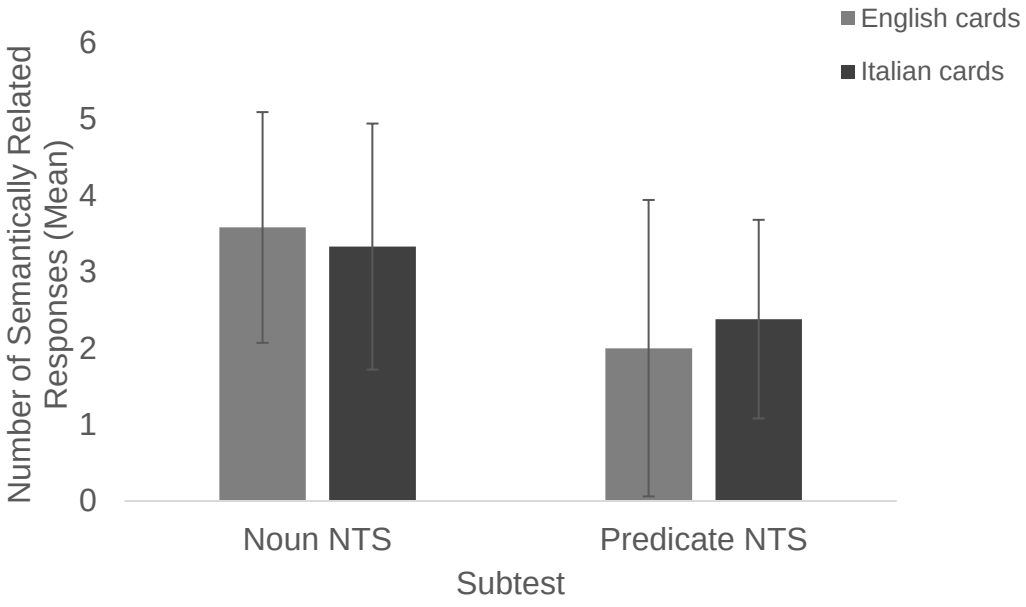

Figure 3: Graph showing the number of non-target but semantically related responses (NTS) elicited by each set of cards in the noun and predicate production subtests.

Distributions of the scores for the English cards and the Italian cards were similar, as assessed by visual inspection. The number of non-target but semantically related errors was not statistically significantly different between the English cards (*Mdn* = 3) and the Italian cards (*Mdn* = 3), $U = 65$, $z = .415$, $p = .713$, using and exact sampling distribution for U (Dineen & Blakesley, 1973).

A Mann-Whitney-U test was run to determine if there were differences in the number of non-target but semantically related errors elicited by the English WinG cards and the Italian WinG cards in the predicate production sub-test. Distributions of the scores for the English cards and the Italian cards were not similar, as assessed by visual inspection. The number of correct responses for the English cards (mean rank = 8.60) and the Italian cards (mean rank = 10.62) were not statistically significantly

different, $U = 31$, $z = -.821$, $p = .460$, using an exact sampling distribution for U (Dineen & Blakesley, 1973).

Individual Cards

As Study One found differences in the ratings of individual cards an independent samples *t*-test was then conducted on the data from Study Two to determine whether there were any significant differences between individual cards in the two sets of WinG cards. The data was coded so that it could be analysed, correct answers were coded as '1' and incorrect answers were coded as '0'.

The independent samples *t*-test showed that there were only significant differences between three individual cards, Mountain in the noun comprehension sub-test, the English card (M=.833, SD=.389) elicited statistically significantly more correct responses than the Italian card (M=.417, SD=.515), M= -0.42, 95% CI [-0.80, -0.03], $t(20.477)$=-2.236, p=.037, Book in the noun production sub-test, the English card (M=.750, SD=.452) elicited statistically significantly more correct answers than the Italian card (M=.333, SD=.492), M= -0.42, 95% CI [-0.82, -0.02], $t(22)$=-2.159, p=.042, and Short in the predicate comprehension sub-test, the English cards (M = 1, SD= .316) elicited significantly more correct responses than the Italian card (M= .625, SD= .518), M= 0.525 , 95% CI [0.066, 0.984], $t(11.04)$= 2.158, p= .029, as detailed in Table 6.

Table 6. Means and standard deviations (SD) of number of correct responses for individual cards which were found to be statistically significantly different.

	English cards		Italian cards	
	Mean	SD	Mean	SD
Mountain	.833	.389	.417	.515
Book	.750	.452	.333	.492
Short	1	.316	.625	.518

Discussion

Study One

The paired samples t-test conducted on the data from Study One demonstrates that the adults rated the English WinG cards as significantly better at representing the constructs than the Italian Wing Cards in the noun comprehension and production subtests as well as the predicate production subtest. Differences were also found between individual cards; this suggests that for these individual target cards, the

English WinG cards appear better than the Italian WinG cards. These findings also indicate that changing the image on a target card may elicit different responses from adult participants.

From these findings we have found some support for our first hypothesis, that there would be a difference in the adult ratings of the two sets of cards, as we found a significant difference in the ratings of three of the four subsections in the WinG test. We can also conclude which direction this difference is in; the English WinG cards were rated as better in three of the four subsections of the WinG test, and every significant difference found between the individual cards favoured the English set.

As ease of identification leads to preference of stimuli (Johnson et al., 1996), these findings suggest that the adults found the images used on the English cards easier to identify than the images on the Italian cards, which therefore lead to them being rated as better representing the construct. It is also somewhat unsurprising that the English adults preferred the English WinG cards to the Italian WinG cards as the targets on the English WinG cards are images from England and so will have been more familiar for English adults than the Italian images.

These findings do support previous research which demonstrates how changing visual stimuli can influence responses given by adults (Wu et al., 2015; Zhang et al., 2007) as the adults rated the two sets of cards as significantly different suggesting that if they were tested using the cards the two sets would elicit different responses, which is similar to the findings of previous research (Wu et al., 2015).

As we found significant results, we can conclude that the adult participants in Study One rated the English cards as better and more representative of the construct than the Italian cards, in three of the four sub-tests. We therefore expected to find support for our second hypothesis, that the children who were tested using the English WinG cards would produce more correct and non-target but semantically related answers than the children who were tested using the Italian WinG cards, in these subtests. Furthermore, as we found significant differences in the ratings of some of the individual target cards, we expected to see differences in the scores between the two groups, particularly on those target cards.

Study Two

In Study Two we did not find any significant differences between the two groups in terms of their scores on the CDI or between genders which suggested that any differences we found would have been due to differences between the cards. We did not find the correlations between the WinG scores and the CDI scores that we were expecting to, however, this could have been due to the incorrect CDI being used for the 24-month-old participants. Overall, we did not find any significant differences between the two sets of cards in any of the subtests so did not find any support for our second hypothesis, even though differences were found in Study One. *Figures 1,*

2, & 3 all show how similarly the children in the two groups performed on the WinG test which demonstrates how reliable both sets of WinG cards are. Furthermore, as no differences were found in the Mann-Whitney U analysis we have demonstrated the reliability of the targets used in both sets of WinG cards when assessing language development in children.

As we found significant differences between individual cards in Study One, we expected to find significant differences in the number of correct scores elicited by those same cards in Study Two. However, only three cards were found to have significant differences, of these three targets only the cards representing Mountain were rated as significantly different by the adults. It is unclear why the cards representing Short elicited significantly different numbers of correct responses as the targets are very similar, however differences between the Mountain and Book cards between the groups could be due to cultural differences. The English card for mountain depicts a snow topped mountain, a concept which is common in English depictions of mountains, whereas the Italian card for mountain did not have snow on top which therefore may not have matched the English children's semantic knowledge of mountains, making it harder for them to answer that question accurately. Furthermore, the cards for book may have elicited significantly different numbers of correct responses as the English card shows an image of a colourful children's book, whereas the Italian card shows a book with writing inside, it is more likely that the English card will be recognisable as a book for young children as they are likely to have encountered similar books in the past compared to the book on the Italian card. This demonstrates the importance of considering the culture that a test is intended to be used in (Shong & Cheng, 2009) as well as using targets that children are likely to be familiar with (Davidoff & Masterson, 1995).

Overall, the results of Study Two suggest that the English WinG cards are reliable as although the adults rated many of them to be significantly better than the Italian cards, we found that when they are used in the WinG test with children there are very few differences between the two sets of cards, both between the individual cards as well as the sub-tests overall. This study also supports that it is appropriate to use pictures in naming tasks with children (Davidoff & Masterson, 1995) as despite changing the pictures the children performed consistently across the two conditions.

General Discussion

Overall, we found some support for our first hypothesis, that there would be a difference in the ratings of the two sets of cards by adult participants, however we did not find support for our second hypothesis: if one set of cards is rated as significantly better than the other we expect that the children tested using the 'better' set of cards will produce more correct answers than the children tested using the other set of cards. These differences could be because the children have fewer expectations of how certain stimuli should look and so are less susceptible to visual

changes, or due to adults and children perceiving stimuli differently (Schum et al., 2012). However it is likely that they are due to differences in the way that the two studies were conducted, the adults were shown both sets and asked for their opinion and so preferred the cards that were more in line with their culture. Whereas, the children were shown only one set of cards and were tested in the way the cards are designed to be used.

With regards to limitations of the current study, it is possible that the different researchers had different interaction styles and therefore spoke to the children in a different way, eliciting different responses, however, the WinG manual (Cattani et al., 2019) outlines phrases which should be used when administering the test therefore limiting the influence of researcher differences. Furthermore, the WinG manual also confirms the answers that should be accepted so there is interrater reliability in this study as different researchers will not accept different answers to be correct. Additionally, all videos of the individual participants were re-watched and any discrepancies in scoring were discussed between the researchers to ensure interrater reliability.

In the future it would be interesting to research whether there is an effect on comprehension and production in children if the WinG cards are made less colourful, it has been shown that colour can influence adult responses (Zhang et al., 2007) but as we found in the current study it seems that children and adults perceive stimuli differently. However, previous research does suggest that children prefer more brightly coloured stimuli (Boyatzis & Varghese, 1994) so it is likely that this would influence the comprehension and production of children to some extent.

Furthermore, it would be interesting for future research to test the comprehension of words that children usually struggle to produce in the WinG test, such as seal and in front, perhaps by swapping the comprehension and production cards around it would be possible to determine whether the children do know these words but cannot produce them yet and whether for some of the comprehension cards such as iron and they are using a process of elimination to point to the correct comprehension card.

Conclusion

In conclusion as we found no significant difference between the two sets of WinG cards in Study Two, we have shown that the English WinG cards are a reliable tool for assessing language development in children. This is because both the Italian WinG cards (Bello et al., 2012) and the English WinG cards (Cattani et al., 2019) elicited very similar responses and there were no significant differences between the scores of the two groups of children, despite the adults rating the cards differently, so there is consistency and reliability between the two tools.

Acknowledgments

I would like to thank the following for their support throughout my research project, my supervisor Allegra Cattani for her continued support and reassurance throughout the project, especially during the difficult circumstances due to the Covid-19 pandemic. The parents who gave up their time to bring their children into the Babylab at the University of Plymouth and participate in our study, this research would not have been possible without them. Thank you to my family who have ensured I have had the space, and peace I needed to write this paper whilst we have all been isolating in the house together, due to the pandemic, and for their support throughout my whole degree. Finally, I cannot thank my partner enough for his ongoing support and calming words throughout this project and the entirety of my time at university.

References

Alcock, K.J. (2017). Production is only Half the Story – First Words in Two East African Languages. *Frontiers in Psychology,* 8, DOI:10.3389/fpsyg.2017.01898.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J., Reilly, J., & Hartung, J. (1994). Developmental and Stylistic Variation in the Composition of Early Vocabulary. *Journal of Child Language,* 21, 85-123.

Bello, A., Giannantoni, P., Pettenati, P., Stefanini, S., & Caselli, C. (2012). Assessing lexicon: validation and developmental data of the Picture Naming Game (PiNG), a new picture naming task for toddlers. *International Journal of Language and Communication Disorders,* 47, 589-602.

Boyatzis, C.J., & Varghese, R. (1994). Children's Emotional Associations with Colors. *The Journal of Genetic Psychology,* 155, 77-85.

Capone, N.C. (2007). Tapping Toddlers' Evolving Semantic Representation via Gesture. *Journal of Speech, Language, and Hearing Research,* 50, 732-745.

Caselli, M.C., Rinaldi, P., Stefanini, S., & Volterra., V. (2012). Early Action and Gesture ''Vocabulary'' and Its Relation with Word Comprehension and Production. *Child Development*, 83(2), 526-542.

Cattani, A., Krott, A., Dennis, I., & Floccia, C. (2019). *WinG Words in Game Test: A vocabulary assessment for pre-school children.* St Mabyn, UK: Stass Publications. ISBN: 9781874534563.

Cattani, A., Floccia, C., Kidd, E., Pettenati, P., Onofrio, D, Volterra, V. (2019). Gestures and Words in Naming: Evidence from Crosslinguistic and Crosscultural Comparison. *Language Learning*, 69, 709-746.

Chomsky, N. (1988). *Language and the problems of knowledge*, Cambridge, MA: MIT Press.

Davidoff, J., & Masterson, J., (1995) The Development of Picture Naming: Differences between Verbs and Nouns. *Journal of Neurolinguistics*, 9, 69–83.

Dineen, L.C., & Blakesley, B.C. (1973). Algorithm AS 62: Generator for the sampling distribution of the Mann-Whitney U statistic. *Applied Statistics*, 22, 269-273.

Duemmler, T., Franz, V.H., Jovanovic, B., & Schwarzer, G. (2008). Effects of the Ebbinghaus illusion on children's perception and grasping. *Experimental Brain Research*, 186, 249–260.

Duff, F.J., Reen, G., Plunkett, K., & Nation, K. (2015). Do Infant Vocabulary Skills Predict School-Age Language and Literacy Outcomes? *Journal of Child Psychology and Psychiatry,* 56(8), 848-856.

Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., et al. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. Baltimore: Brookes.

Goldin-Meadow, S. (2000). Beyond Words: The Importance of Gesture to Researchers and Learners. *Child Development*, 71, 231-239.

Goldin-Meadow, S., Seligman, M.E.P., & Gelman, R. (1976). Language in the two-year old. *Cognition,* 4, 189–202.

Johnson, C.J., Paivio, A., & Clark, J.M. (1996). Cognitive Components of Picture Naming. *Psychological Bulletin,* 120(1), 113-139.

Kraynak, A.R., & Raskin, L.M. (1971). The Influence of Age and Stimulus Dimensionality on Form Perception by Preschool Children*. Developmental Psychology*, 4(3), 389-393.

Norbury, C.F. (2015). Editorial: Early intervention in response to language delays – is there a danger of putting too many eggs in the wrong basket? *Journal of Child Psychology and Psychiatry*, 56, 835-836.

Schum, N., Franz, V.H., Jovanovic, B., & Schwarzer, G. (2012). Object Processing in Visual Perception and Action in Children and Adults. *Journal of Experimental Child Psychology*, 112(2), 161-177.

Shong, S.Y.L., & Cheng, S.T. (2009). Development of a Screening Instrument for Early Language Delay in Hong Kong Chinese: A Preliminary Study. *The Journal of Genetic Psychology: Research and Theory on Human Development,* 170, 193-196.

Walker, M.M., Barrow, I., & Rastatter, M.P. (2002). The Effect of Dimension and Vocabulary Age on Rapid Picture Naming in Children. *Journal of Communication Disorders,* 35(1), 1-10.

Wu, H-K., Kuo, C-Y., Jen, T-H., Hsu, Y-S. (2015). What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers & Education*, 85, 35-48.