**Method**

Participants

Pilot Study 1

Participants in Pilot Study 1 were 5 early bilinguals of Spanish and Catalan (3 female, 2 male). Four of the participants (Participants 2-5) were simultaneous Spanish-Catalan bilinguals, who began learning Spanish and Catalan before age 2. The fifth participant (Participant 1) was a sequential bilingual, who began speaking Catalan at the age of 7 when her family moved from Colombia, and therefore was factored into analyses as a separate profile. The participants were recruited from the Universitat Pompeu Fabra in Barcelona, Spain. Participants had completed between 1 and 4 years of university study. The average age was 20.6 (SD: 1.34).

The participants were asked to self-report their abilities in Spanish and Catalan on a scale of 0-5, where 0 was defined as “poor” and 5 as “near-native or native”. They ranked their abilities in the following five categories: reading, writing, speaking, listening and vocabulary knowledge. The four simultaneous bilinguals self-reported each of these abilities in Spanish and Catalan at near-native or native levels (M = 5, SD = 0); the sequential bilingual reported her Catalan at “near-native or native” for reading, and at “excellent” for the other four abilities. However, she also rated her Spanish listening and speaking abilities as “excellent”, and only ranked her reading, writing and vocabulary knowledge at the “near-native or native” level.

Participants were also experienced foreign language learners. The four simultaneous bilinguals reported the age at which they began learning English before age 13 (M = 7.5, SD = 3.27). The sequential bilingual reported learning English between 13 and 17. Their reported English abilities are listed in TABLE XXXXXX. Three participants (2, 3, 4) also began learning a fourth language after age 5 before age 13 (Japanese, German, French, respectively).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 | 3 | 3.2 | 2.8 | 2.4 |
| 1 | 1 | 1.483239697 | 1.095445115 | 1.140175425 |

Pilot Study 2

Participants in Pilot Study 2 were 18 classroom learners of Spanish (7 female, 11 male). They were recruited from one of three Spanish language courses at Georgetown University (Intermediate I [3 participants], Advanced I [12], or Oral Review [3], a course that follows Advanced II for these students). Three early bilinguals (2 Korean-English and 1 English-Spanish bilinguals) were removed from the sample. The remaining 15 participants all identified themselves as native speakers of English only, all born in the United States, and with a mean age of 18.6 years (SD = 0.74). One participant began learning Spanish before age 10, attending a bilingual English-Spanish pre-school and kindergarten. Seven participants began learning Spanish between the ages of 10 and 13, and seven between 13 and 18. Spanish was the first classroom-learned language for 13 of these participants; one participant studied Hebrew and one studied French after age 13.

In addition to low accrual numbers in the Intermediate I and Oral Review groups, two intermediate students did not attain the minimum comprehension threshold (discussed below) and one Oral Review student was removed from the sample as an early bilingual, so analyses discussed beyond this point will only include the ten remaining advanced students.

**Materials**

Experimental Task

The design of the experimental task included two sets of stimuli with a [+/-conflict] dichotomy: the non-linguistic flanker task and the linguistic self-paced reading stimuli. The critical stimuli were interleaved such that each critical linguistic stimulus was preceded by a congruent or incongruent flanker trial.

*Non-linguistic Stimuli*

The flanker task asks participants to respond by pressing one of two buttons, according to the direction indicated by the center arrow on the screen in each trial. Each stimulus presents either one or five arrows in the center of the screen. Stimuli fall into one of three conditions: single center arrow, congruent center arrow, and incongruent center arrow. The congruent condition includes five arrows all pointing in the same direction, while the incongruent condition includes one center arrow pointing in one direction, flanked by two arrows on each side that point in the opposite direction of the center arrow. The critical comparison is between the congruent and incongruent stimuli. The direction of the center arrow is balanced across trials. Each stimulus is preceded by a fixation cross on screen for 1000 ms. Participants have up to 2000 ms to respond. Failure to respond within that time limit is scored as in an incorrect response.

*Linguistic Stimuli*

The linguistic stimuli were sentences presented via the self-paced reading moving window paradigm. Before each sentence was presented, a fixation cross lasting 1000 ms reset the participant’s gaze on the center of the screen to prepare the participant for the trial. Each stimulus consisted of a single sentence with each word masked by hyphens, presented on screen in its entirety. Participants uncovered one word at a time by clicking the space bar. As one word was uncovered, the previous word was re-covered by hyphens. The latency of reading times of each word was recorded by PsychoPy. Following each sentence, participants answered a true/false comprehension question.

The critical linguistic stimuli presented in the self-paced reading trials presented subject-object ambiguities, all of which are listed in the Appendix. These stimuli modulated transitivity of the verb in a preposed adverbial clause, as can be seen in (XXXXX) below (with each region of interest indicated by a backslash).

(XXXX) a. *Transitive*: Cuando el escultor \ acabó \ la obra \ tenía tres metros \ de altura.

b. *Intransitive*: Cuando el escultor \ volvió \ la obra \ tenía tres metros \ de altura.

Region: 1 \ 2 \ 3 \ 4 \ 5

“When the sculptor\ {finished/came back} \ the piece \ was ten feet \ in height.”

*Experimental Design*

Other than minor changes to accommodate a lower proficiency level (addressed below), the transitive (ambiguous) sentences were taken directly from Jegerski (2012). However, because the preceding flanker trial introduced an additional variable, the design departed from Jegerski’s (2012) original design. Two lists of flanker-sentence stimulus pairs were created in a Latin Square design. Each list was then divided into two blocks to counterbalance presentational order across participants. The 20 intransitive (unambiguous) sentences were distributed across these four blocks (five unambiguous sentences per block), always preceded by an incongruent flanker trial. Thus, each participant would complete one list, consisting of two blocks with a total of 10 flanker-sentence pairs of the [+conflict, +ambiguity] condition, 10 pairs of the [-conflict, +ambiguity] condition and 10 pairs of the [+conflict, -ambiguity] condition.[[1]](#footnote-1) Because the change to Jegerski’s (2012) original design meant participants would see both the transitive and the intransitive sentence, minimal lexical changes were made to the intransitive sentences in regions 4 and 5 so that participants were not reading the same sentences twice, but replacement words were chosen according to number of syllables and frequency.

In addition to the 30 critical flanker-sentence pairs addressed above, participants also completed 90 distractor flanker-sentence pairs and 30 unpaired distractor flankers to disrupt the participants expected patterns. The sentential distractor items included preposed adverbial phrases that included a transitive verb with an explicit complement, permanently ambiguous sentences and sequential prepositional phrases with high-low attachment ambiguity. The flanker distractor items were chosen such that each of the six conditions (right vs. left; congruent vs. incongruent vs single) was balanced across all trials.

Frequency of the words in the linguistic stimuli and distractors was assessed using the *Corpus del Español*’s WordAndPhrase tool, which compiles 20 million words from texts from the 1900s and 2 billion words from texts published online since 2014. Words in the stimuli that were not among the 3000 most frequent words on the corpus’ list were included in a decontextualized vocabulary test given to Beginner 1 students to assess how well more advanced students would recognize these words. The students in Beginner 1 were asked to translate each word and rate how well they recognized the word on a scale of 1-4. If words in this test met three of the following four criteria, they were included in the critical stimuli: (i) the word is covered in the vocabulary lists of Beginner 1 or 2 (i.e. Intermediate I students are expected to know this word); (ii) the word elicited an exact translation accuracy of 85% or higher; (iii) the word received a correct recognition rate among all raters of at least 3; and (iv) the word is a cognate with no more than two graphemic changes due to Spanish phonotactics and spelling norms (e.g. violín, escultor). The seventeen words that did not meet these criteria were included in a vocabulary training, which contained a total of 42 words. This training is addressed below.

COMPREHENSION QUESTIONS

Vocabulary Training

Study 2 participants were all required to complete a vocabulary training on PsychoPy before completing the study to establish a baseline vocabulary. This training presented 42 vocabulary items used in the stimuli, 17 of which appeared in the critical stimuli. The 42 words were divided into two blocks. The words in each set were presented one at a time with a photo that represented the word and a translation in English. After 1000 ms, participants were invited to move to the next word if and when they were ready. After each set of 21 words was presented, participants had an opportunity to review the words again or take a quiz to continue. The quiz presented four of the photos from the set and asked participants to identify which photo represented the word shown on the screen. A minimum score of 85% was required on each quiz in order to continue to the next phase of the study.

Language Background Questionnaire

All participants completed a language background questionnaire (LBQ), edited from the LEAP-Q (Marian, Blumenfeld & Kaushanskaya, 2007). Information regarding languages spoken was collected for all participants, but the participants in Study 1 completed detailed questions for Spanish, Catalan and English; participants in Study 2 completed detailed questions for English and Spanish. The questions were presented via computer.

Elicited Imitation Task

To complement institutional placement in Study 2, the Elicited Imitation Task (EIT) was used as a secondary proficiency measure (SEE APPENDIX XXXXX). The EIT has been validated as a reliable measure of global proficiency as compared to the Simulated Oral Proficiency Interview (SOPI), a global proficiency measure based on the American Council of the Teaching of Foreign Languages proficiency guidelines (Ortega, 2000; Ortega et al., 2002). In this task, participants are recorded listening to and repeating a set of 30 pre-recorded sentences. These sentences become progressively more complex structurally. The recordings are then scored on a scale of 0-4 according to the accuracy of each repetition, and the total score for each participant falls between 0 and 120 (see APPENDIX XXXXXX).

Items changed from Jegerski:

* Después de que empezó el maratón/después de que corrió el maratón 🡪 “empezar” now “llegar”
* acabar la obra 🡪 terminar la obra
* 3 metros de altura 🡪 diez pies de altura
* Cinco días después 🡪 antes de que
* Concursante 🡪 participante
* Cuando la novia descendió la escalera le pareció muy larga 🡪 Cuando la chica lava los platos se quedan bien sucios
* Limpiecita 🡪 muy limpia
* De una vez 🡪 inmediatamente

1Counted, “separately”

2Counted

3Counted

4Counted

5Counted, flankers very slow

6 counted

7 out – Daniel Arenas, native speaker

8 out – Stephen Cho, Korean, flankers residuals = negative

9 counted

10 counted

11 – oral review

12 - counted

13 - counted

14 - counted

15 - counted

16 – advanced, did not meet comprehension threshold

17 - counted

18 – oral review

19 – intermediate (60%)

20 – intermediate (only remaining intermediate)

21 – intermediate (50%)

22 – data lost, oral review, Korean speaker

23 - counted

|  |
| --- |
| 12.88888889 |
| 2.643279801 |

Procedure

Study 1

Early bilinguals in Study 1 were recruited from a group of students at the *Universitat Pompeu Fabra* who participated in a Spanish conversation program with Georgetown’s summer abroad program in Barcelona. They were paid 10€ for completing the study. The entirety of the study was conducted in Spanish. Before beginning the study, the researcher explained the study broadly in a quiet room, obtained informed consent and assigned participants to one of four groups. These groups determined the list and presentational order of blocks for each participant. Following this, participants completed three training phases to prepare for the experimental phase: isolated flankers, isolated sentence readings, and interleaved flanker-sentence reading pairs. Participants needed to achieve minimally 80% accuracy in order to move on to the experimental phase. Once they completed the training phases, they had an opportunity to rest and ask any questions before beginning their two experimental blocks, each of which contained 60 flanker-sentence pairs, 15 unpaired flankers, and a brief break. Following the first block, participants had another opportunity to rest before completing the second block. In addition, if the participant answered three consecutive flankers or three consecutive comprehension questions incorrectly, they were forced to take a 10-second break, during which they were reminded that both speed and accuracy were important to finish the study in a timely fashion. Upon completion of the experimental component of the study, participants completed the LBQ.

Study 2

Emergent bilingual language learners in Study 2 were recruited from Spanish language classes at Georgetown University. They were rewarded 2 points on a mid-semester test for participating in the study. Participants completed the study in a large computer lab with up to three other participants. These participants all wore headphones and were sat in opposite corners of the lab facing away from each other to minimize interference. Before beginning the study, the researcher explained the study broadly, obtained informed consent and assigned participants to one of the four groups to determine list and block order. Participants completed three training phases to prepare for the experimental phase: isolated flankers, isolated sentence readings, and interleaved flanker-sentence reading pairs. The explanation of the study, informed consent, and first two training phases were conducted in English. The instructions to the interleaved training phase were presented in English, but then participants were instructed that the actual stimuli and comprehension questions would be presented in Spanish. This was done to facilitate the transition to Spanish and initiate the so-called “monolingual mode” in Spanish. All computerized instructions and stimuli following this point were presented in Spanish. In order to move onto the experimental phase, participants needed to achieve 80% accuracy or higher on each practice phase.

Once they completed the training phases, they had an opportunity to rest and ask any questions before beginning their two experimental blocks (identical to the blocks in Study 1). Like in Study 1, participants who answered incorrectly on three consecutive flankers or comprehension questions were forced to take a short break. Upon completion, participants completed the LBQ, after which they were escorted from the laboratory to complete the EIT, the final component of Study 2, in another space to control the noise level in the laboratory.

Scoring & Coding

*Self-paced Reading*

PsychoPy latency data collection occurs at the word level, so regional-level latency was calculated for 40 critical stimuli and 40 non-critical stimuli (coded for future studies).

*Outliers*

In order to account for data that was likely affected by external factors, data was trimmed at the group and individual level. At the group level, a maximum of 6000 ms (Jackson, 2010 – this is from JEGERSKI’s book) and a minimum of 100 ms (Luce, 1986 ---Jegerski’s book) was set. At the individual level, maximums and minimums were established for each participant at +2.5 standard deviations (SDs) from the mean RT (Marijuan, dissertation). If the individual limits were outside the bounds for the group (i.e., less than 100 ms or greater than 6000 ms), the group values were used for trimming and replacing outliers. In no case was a RT replaced with the minimum limit. Data that exceeded the upper bound accounted for 2.54% of the total data. This falls within the standard recommended by Jegerski (2014).

|  |  |
| --- | --- |
| 168(total removed) | 6600(total answered) |
| 0.02545455 |  |

*Comprehension Check*

After removing outliers, the data was then cleaned with regards to comprehension checks. Participants with comprehension accuracy below 60% were removed from analyses (Marijuan, dissertation). This resulted in the removal of two intermediate participants and one advanced participant. The average comprehension rate of the remaining participants (5 early bilinguals in Study 1; 9 emergent bilinguals in Study 2) was 92.13% and 76.41% respectively.

At the item level, sentence-flanker pairs with incorrect comprehension questions were removed from analyses. This accounted for the removal of 8.0% of early bilingual data (12 stimuli pairs distributed across the 5 participants) and 24.4% of emergent bilingual data (66 stimuli pairs across the 9 participants).

*Residual Reaction Times*

At this point, residual RTs (RRTs) were calculated to normalize the data, using the RTs of the sentences that the participants comprehended and that complied with the above established criteria. This process requires plotting coordinate pairs of raw RTs and the length of the corresponding regions. The line of best fit is then calculated for each participant. The slope of this line reflects the average reading time at different word lengths, and the y-intercept represents the baseline reading time for each word, once differences of length are factored out. This information is used to generate an expected RT for a region based on the number of letters in that region. The expected RT is subtracted from the actual reading time, resulting in the RRT. Positive RRTs reflect a region read slower than expected; negative RRTs reflect regions read faster than expected. The aggregate RRT means for each analogous region were then calculated.

*EIT Scoring*

The elicited imitation test includes 30 sentences of increasing length and complexity, each of which is scored from 0 to 4 (Ortega, 2000). A research assistant scored the recordings of the EITs. Before scoring the participants from this study, the researcher and research assistant independently scored two EITs collected for a previous study. After this first round, Cohen’s kappa (κ) was run to determine the extent of agreement between the two coders. Following recommendations from Cohen (1960) and Lowry (2019), linear weighting was used to account for the fact that the scores awarded were ordinal (increasing scores represent increasingly more successful production of the target), and the reported kappa is observed as a proportion of the maximum possible kappa given the marginal frequencies (this is recommended in cases when raters are not limited to a certain number of ratings per category). Concordance was moderate between the two raters, κlinear weighted = 0.7024 (95% CI, 0.196 to 0.521), SE = 0.0622. Disagreements were discussed one by one until a consensus was reached. After this, the two raters individually scored recordings from the first two participants from Study 2. Cohen’s Kappa was calculated again and a much higher rate of agreement was found, κlinear weighted = 0.9472 (95% CI, 0.8091 to 0.9463), SE = 0.035. This is considered excellent by Fleiss (1981) and almost perfect agreement by Landis and Koch (1977). Following the second round of independent scoring, the research assistant scored the remainder of Study 2 participants’ recordings.

Observed: .8777

SE: .035

LL: .8091

UL: .9463

Max possible: .9266

Observed as proportion of max: .9472 (because no limit to how many 1s and 2s and 3s (etc) each rater could give

The two researchers discussed each disagreement.

First round

Observed Kappa Unweighted:

.3587, SE: .0828 (Lower limit = .1964, upper limit = .521)

Kappa with llinear weighting:

Observed: .5742, SE.0662, lower: .444, upper: .704

Maximum possible linear-weighted kappa, given the observed marginal frequencies: .8175

.7024 = kappa observed as a proportion of maximum possible (e.g. .5742/.8175)

from participants two previously coded EIT recordings,he scoring of the EITs Before scoring the EIT recordings, The EIT recordings of all participants in Study 2 were

EIT FROM ARIEL

The test produces a score on a scale from 0 to 120 (see Appendix D for scoring protocol). Due to the subjective nature of the scoring protocol, the EIT was coded by three independent raters. To ensure interrater reliability, first, each rater scored the responses from the same three EITs using the scoring protocol from Ortega (2000) (see Appendix D). The raters then compared their scores and, where discrepancies arose, arguments for choosing the scores were discussed, scoring was clarified, and a consensus was made as to the most appropriate score. Based on the results from this process, three more EITs were chosen and each rater scored them. Once again, results were compared and interrater reliability was calculated by totaling the number of items where discrepancy arose divided by the total number of items, reaching an interrater reliability score of 86.67%. Table 3 presents the descriptive statistics for EIT scores for the four initial recruitment groups, second-semester, fourth-semester, sixth-semester, and advanced. Three participants, all from second-semester, were removed from the study due to lack of an EIT score after their answers failed to be recorded, one participant was removed from the advanced group for being a heritage learner, and one participant was removed from the sixth-semester group as she reported a language other than English as her L1 and did not begin to study English until high school. This yielded a total L2 learner group of 90 participants. A one-way ANOVA and subsequent post hoc Scheffé showed a significant difference between all four groups on the EIT, F(3,83) = 84.71, p = .000 (η2 = .76, observed power = 1.00), with a large effect size between groups15 . Due to the overlap between the large ranges of scores for L2 groups, participants from the original recruitment were redistributed into four new groups based on percentile scores for the EIT. Four evenly distributed groups of EIT scores were established based on the 25th (n=22, range 10-45), 50th (n=23, range 46-83), 75th (n=23, range 84-111), and 100th (n=22, range 112- 120) percentiles. A one-way ANOVA showed that there was a significant difference by percentile group, F(3,83)=393.24, p=.000, partial ŋ2 =0.93, with a large effect size, and the post hoc Scheffé test indicated that all four percentile groups performed significantly different. In order to create a clear distinction in proficiency between each group, group boundaries were established so that the highest score of one group was not within one point of the lowest score of the next group, using the percentile group boundaries as a starting point. This yielded the range of proficiency scores for the four L2 learner groups presented in Table 4. A one-way ANOVA found a significant effect by group for the EIT scores, F(4,119) = 637.57, p=.000, partial ŋ2 =0.95, with a large effect size. A post hoc Scheffé test indicated that all four participant groups are significantly different from one another, but the high advanced group was not significantly different from total possible score, as indicated with indices in Table 4, where non-matching indices indicate significant differences at p<0.05.

http://vassarstats.net/kappa.html

©Richard Lowry 2001-2019  
All rights reserved.

Second round Kappas

Unweighted:

Observed: .7817

SE: .063

Lower limit: .6582

Upper limit: .9052

.9345 = max possible unweightd, given observed marginal frequencies

.8365 = observed as proportion of max possible

Linear weighted (because 0 is less perfect than 1 is less perfect than 2 is less perfect than 3… so a pair of 1 and 2 is closer than a pair of 1 and 5… weighted kappa takes proximity of mismatched ratings into account in addition to perfect coincidence of rating)

Observed: .8777

SE: .035

LL: .8091

UL: .9463

Max possible: .9266

Observed as proportion of max: .9472 (because no limit to how many 1s and 2s and 3s (etc) each rater could give

Kappa provides a measure of the degree to which two judges, A and B, concur in their respective sortings of N items into k mutually exclusive categories. A 'judge' in this context can be an individual human being, a set of individuals who sort the N items collectively, or some non-human agency, such as a computer program or diagnostic test, that performs a sorting on the basis of specified criteria. [Click [here](javascript:display()) for an explanation of the conceptual and computational details of kappa.]

To begin, select the number of categories by clicking the appropriate button below; then enter your data into the appropriate cells of the data-entry matrix. After all data have been entered, click the «Calculate» button. To perform a new analysis, click the «Reset» button and start over. The analysis assumes that each entered value is an integer equal to or greater than zero.T

Note that measures of weighted kappa are meaningful only if the categories are ordinal and if the weightings ascribed to the categories faithfully reflect the reality of the situation. The weightings in this case are determined by the imputed relative distances between successive ordinal categories. By default, each of these distances is set at '1'. You are free to change any or all of these distances, though I recommend you do so only if you have good reason for it.

The author is grateful to César Roberto de Souza for detecting an error in the original programming for this module and suggesting the appropriate correction.

|  |
| --- |
| Confidence intervals for proportions are calculated according to the Wilson efficient-score method, corrected for continuity. |

http://vassarstats.net/small_logo.gif

**Kappa**

Kappa provides a measure of the degree to which two judges concur in their respective sortings of N items into k mutually exclusive categories. A 'judge' in this context can be an individual human being, a set of individuals who sort the N items collectively, or some non-human agency, such as a computer program or diagnostic test, that performs a sorting on the basis of specified criteria.

**Simple Unweighted Kappa**T  
The original and simplest version of kappa is the unweighted kappa coefficient introduced by J. Cohen in 1960. To illustrate, suppose that our judges are two clinical tests, A and B, independently employed to sort each of N=100 subjects into one or the other of k=3 diagnostic categories. The table on the left shows a cross-tabulation of the sortings actually observed, while the one on the right shows the cell frequencies that would have been expected by mere chance, given the observed marginal totals. In both tables the cells representing concordance of the two tests are highlighted in bold red. 

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (i) **Observed** | | | | | |  | (ii) **Chance Expected** | | | | | |
|  | | **B** | | | **Total** |  | | **B** | | | **Total** |
| **1** | **2** | **3** | **1** | **2** | **3** |
| **A** | **1** | **44** | 5 | 1 | 50 | **A** | **1** | **30** | 15 | 5 | 50 |
| **2** | 7 | **20** | 3 | 30 | **2** | 18 | **9** | 3 | 30 |
| **3** | 9 | 5 | **6** | 20 | **3** | 12 | 6 | **2** | 20 |
| **Total** | | 60 | 30 | 10 | 100 | **Total** | | 60 | 30 | 10 | 100 |
| Observed Concordant Items:  count = 70   proportion = .70 | | | | | | Expected Concordant Items:  count = 41   proportion = .41 | | | | | |

In this example the observed number of concordant items is 70, the chance expected number is 41, and the excess of observed over expected is 70−41=29. Similarly, the chance expected number of non- concordant items is 100−41=59. Cohen's kappa is simply the ratio of the former to the latter: 29/59=.4915.Essentially, what it says is: Of all the items that we would have expected to be non-concordant if nothing more than chance coincidence were operating in the situation, 49.15% of them are in fact concordant.   
  
In the uncommon case where Judges A and B start out by sorting the same number of items into category 1, the same number into category 2, and so forth, the upper limit of kappa is 1.0. When the judges start out with different numbers of items in any one or several of the categories, the maximum possible value of kappa will be less than 1.0. In the present example the maximum possible degree of concordance, given the observed marginal totals,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (iii) **Maximum Possible** | | | | | |
|  | | **B** | | | **Total** |
| **1** | **2** | **3** |
| **A** | **1** | **50** | 0 | 0 | 50 |
| **2** | 0 | **30** | 0 | 30 |
| **3** | 10 | 0 | **10** | 20 |
| **Total** | | 60 | 30 | 10 | 100 |

would be produced by the sorting shown in the adjacent table, which would yield kappa=.8305. The ratio of our observed kappa to this maximum possible value is .4915/.8305=.5918: which is to say that the observed value is 59.18% as large as it possibly could be, under the circumstances. 

**Weighted Kappa**T  
When the categories are merely nominal, Cohen's simple unweighted coefficient is the only form of kappa that can meaningfully be used. If the categories are ordinal—if it is the case that category 2 represents *more* of something than category 1, that category 3 represents *more* of that same something than category 2, and so on—then it is potentially meaningful to take into account not only the absolute concordances (the ones shown in the bold-face red font in the example), but also the relative concordances (as now shown in the black font). In taking these relative concordances into account, each cell in a row of the matrix is weighted in accordance with how near it is to the cell in that row that includes the absolutely concordant items. 

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (i) **Observed** | | | | | |  | (ii) **Chance Expected** | | | | | |
|  | | **B** | | | **Total** |  | | **B** | | | **Total** |
| **1** | **2** | **3** | **1** | **2** | **3** |
| **A** | **1** | **44** | 5 | 1 | 50 | **A** | **1** | **30** | 15 | 5 | 50 |
| **2** | 7 | **20** | 3 | 30 | **2** | 18 | **9** | 3 | 30 |
| **3** | 9 | 5 | **6** | 20 | **3** | 12 | 6 | **2** | 20 |
| **Total** | | 60 | 30 | 10 | 100 | **Total** | | 60 | 30 | 10 | 100 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observed Frequencies | | | | |
|  | | **B** | | |
| **1** | **2** | **3** |
| **A** | **1** | **44** | 5 | 1 |

To illustrate the weighting process, consider the first row in the above table of observed frequencies. Suppose we had good reason to assume that the distance between categories 1 and 2 is about the same as the distance between categories 2 and 3. In this case, cell A1B2 would lie at a distance of one (relative) unit from cell A1B1 (A1B1 is the cell in row 1 that marks absolute concordance) and cell A1B3 would fall at a distance of two (relative) units from cell A1B1. The same principle would apply to the other rows in the matrix. ThusT 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Distances | | | | |
|  | | **B** | | |
| **1** | **2** | **3** |
| **A** | **1** | **0** | 1 | 2 |
| **2** | 1 | **0** | 1 |
| **3** | 2 | 1 | **0** |

With k ordinal categories and equal imputed distances between successive categories, the maximum possible distance between any two categories is k−1, which in the present example is equal to 2. The weights derived from these imputed distances can be either linear or quadratic:   
  
If they are linear, then the weight for any particular cell isT

|  |  |  |
| --- | --- | --- |
|  | weight = 1− | |distance|  maximum possible distance |

And if they are quadratic, then the weight for a particular cell isT

|  |  |  |
| --- | --- | --- |
|  | weight = 1− | (distance)2  (maximum possible distance)2 |

With the present example, this would yield the following sets of weights for the cells:T

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Linear | | | | |  | Quadratic | | | | |
|  | | **B** | | |  | | **B** | | |
| **1** | **2** | **3** | **1** | **2** | **3** |
| **A** | **1** | **1** | .5 | 0 |  | **A** | **1** | **1** | .75 | 0 |
| **2** | .5 | **1** | .5 | **2** | .75 | **1** | .75 |
| **3** | 0 | .5 | **1** | **3** | 0 | .75 | **1** |

To give you an idea of how a weighted kappa coefficient is calculated, I show again the data of our example, only now each of the frequency values has been divided by N (in this case, N=100)to convert it into a proportion. 

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Proportions** | | | | | | | | | | | | |
| (i) **Observed** | | | | | |  | (ii) **Chance Expected** | | | | | |
|  | | **B** | | | **Total** |  | | **B** | | | **Total** |
| **1** | **2** | **3** | **1** | **2** | **3** |
| **A** | **1** | .44 | .05 | .01 | .50 | **A** | **1** | .30 | .15 | .05 | .50 |
| **2** | .07 | .20 | .03 | .30 | **2** | .18 | .09 | .03 | .30 |
| **3** | .09 | .05 | .06 | 20 | **3** | .12 | .06 | .02 | .20 |
| **Total** | | .60 | .30 | .10 | 1.00 | **Total** | | .60 | .30 | .10 | 1.00 |

For each of the gray cells in the "Observed" table, multiply the proportion by the linear weight corresponding to that cell, and sum the results across all nine of the cells. This sum will be   
  
     Pobserved = .8   
  
Performing the same operation for the nine gray cells in the "Chance Expected" table will yield   
  
     Pexpected = .62   
  
The kappa coefficient with linear weighting is then simply the ratio 

|  |  |  |
| --- | --- | --- |
|  | kappaLW = | Pobserved − Pexpected  1 − Pexpected |
|  |  |  |
|  | = | .8 − .62  1 − .62 | = .4737 |

Performing this same procedure with the quadratic weights would yield kappaQW=.4545.   
  
Weighted kappa coefficients are less accessible to intuitive understanding than is the simple unweighted coefficient, and they are accordingly more difficult to interpret. References are listed below for those who might wish to pursue the matter further.

**Proportions of Agreement**T  
Independently of kappa, it is also possible to measure the proportion of agreement between the two judges within each of the k categories separately. In the example (Table i), Judges A and B agreed on a total of 44 items for category 1. For Judge A there were 6 additional items in category 1 with which B did not agree, while for Judge B there were 16 additional items in category 1 with which A did not agree. The proportion of agreement for category 1 is therefore 44/(44+6+16)=.6667. Similarly, the proportion of category 1 agreement to be expected by mere chance (Table ii) is 30/(30+20+30)=.375; and the maximum possible proportion of agreement (Table iii), given the observed marginal totals, is 50/(50+0+10)=.8333. 

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (i) **Observed** | | | | | |  | (ii) **Chance Expected** | | | | | |
|  | | **B** | | | **Total** |  | | **B** | | | **Total** |
| **1** | **2** | **3** | **1** | **2** | **3** |
| **A** | **1** | **44** | 5 | 1 | 50 | **A** | **1** | **30** | 15 | 5 | 50 |
| **2** | 7 | **20** | 3 | 30 | **2** | 18 | **9** | 3 | 30 |
| **3** | 9 | 5 | **6** | 20 | **3** | 12 | 6 | **2** | 20 |
| **Total** | | 60 | 30 | 10 | 100 | **Total** | | 60 | 30 | 10 | 100 |
| Observed Concordant Items:  count = 70   proportion = .70 | | | | | | Expected Concordant Items:  count = 41   proportion = .41 | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (iii) **Maximum Possible** | | | | | |
|  | | **B** | | | **Total** |
| **1** | **2** | **3** |
| **A** | **1** | **50** | 0 | 0 | 50 |
| **2** | 0 | **30** | 0 | 30 |
| **3** | 10 | 0 | **10** | 20 |
| **Total** | | 60 | 30 | 10 | 100 |

References

Agresti, A. (1996), *An Introduction to Categorical Data Analysis,* New York: Wiley   
  
Cicchetti, D.V. and Allison, T. (1971), "A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings," *American Journal of EEG Technology,* 11, 101-109   
  
Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement,* 20, 37-46.   
  
Cohen, J. (1968), "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin,* 70, 213-220.   
  
Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969), "Large-Sample Standard Errors of Kappa and Weighted Kappa," *Psychological Bulletin,*72, 323-327.   
  
Fleiss, J.L. and Cohen, J. (1973), "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and Psychological Measurement,* 33, 613-619.   
  
Fleiss, J.L. (1981), *Statistical Methods for Rates and Proportions,* 2nd Edition, New York: Wiley

Results & Analyses

Discussion & Conclusion

**Bibliography**

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, 20, 37-46.

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley. [ISBN](https://en.m.wikipedia.org/wiki/International_Standard_Book_Number) [0-471-26370-2](https://en.m.wikipedia.org/wiki/Special:BookSources/0-471-26370-2).

Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics. 33 (1): 159–174.

Lowry (Vassarstats – <http://vassarstats.net/>)

1. Proposed Change: The decision to present all unambiguous stimuli after a [+conflict] flanker was made to allow for stronger comparisons and to avoid attention to the similar sentence structures between the transitive and intransitive sentences, but following data collection, it became clear that a stronger study would also include the fourth permutation of the conditions: [-conflict, -ambiguity]. The proposed change would mean both lists would include each unambiguous sentence, and the preceding flanker condition would be counterbalanced across the two lists. [↑](#footnote-ref-1)