Communication and Learning Degrees

Seth Kirby

Indiana University

Abstract

We explore the concept of learning and how information passes through successive generations of teachers and learners. Specifically we set out to analyze the relationship between high and low scoring learners on connected word definitions in a directed network. Performance on vocabulary synonym matching was used to score successive generations of learners, and each generation of learners created new definitions to train successive generations.

Method

Subjects were evaluated using a generated vocabulary test consisting of sixty questions evaluating performance across ten words, with four additional questions for verification of response truthfulness and effort. Each word that appeared in the test appeared six times, and was paired with six different synonyms. Synonym ordering, word pairing, and question ordering were randomized, though words always appeared three times before and three times after the teaching phase where word definitions were presented.

The words were selected from a publicly available word frequency list culled from the Corpus of Contemporary American English (Word Frequency Data). Ten words were selected from positions 30,000 to 50,000 of the 60,000 word list, which was ordered by descending level of common use. The candidate word list was randomly ordered via Python's random.shuffle(), and then filtered manually to remove words which represented alternate tenses. Synonyms were then gathered for candidate words using bighugelabs thesaurus API (Big Huge Thesaurus); all candidates with less than six synonyms were removed from the candidate set.

Definitions were created in successive rounds by previous learners, starting from the source dictionary definitions, our degree zero definitions. Definitions for degree zero were taken from Merriam-Webster's Learner's Dictionary, via their API (Dictionary API). Learners given these definitions during training created new definitions at the end of their test, which become degree one definitions for further testing. These definitions feed forward to new learners to create degree two definitions, while those fed forward to create degree three definitions. Definitions for successive degrees are generated from previous learners, linking the scoring of

previous learners to their created definitions, creating a network of connected edges from a graph perspective.

In each test, a tested word was paired with three randomly selected words from the remaining set, and this grouping of four words remained consistent through the six questions of the tested word. If word one is paired with words two, five and nine, every question on word one will be paired with the synonym set of words one, two, five and nine. This strategy was selected to limit the effects of test taking strategies based on a process of narrowing the correct answer to the synonym set which appears most often for a particular word over time. If participants were able to take advantage of this strategy, we would see increased correctness through each of a words six questions. As the training stage always appears at the midpoint of the test, this would skew the results towards training effectiveness, especially as we employed a binary ordering of question progression, representing only whether a word appeared before or after the training stage.

Each test consisted of sixty test questions, four effort validating questions, eight shown definitions, and a demographics section. The words were grouped into question categories, two words each were tested when given definitions from degree zero, degree one, degree two, degree three, and no definition. The first section of the test evaluated performance in a naïve and untrained state, and consisted of three questions for each word, and two truth questions. The next section of the test consisted of definitions for non-control words, shown two per page. These definitions were selected randomly from the word's definition group at the correct degree. Generated questions were ordered randomly, with the exception that naïve and trained state questions were necessarily divided by the training stage.

Testing was conducted using Amazon Mechanical Turk, and was divided into two phases. In the first phase, definitions for each degree were generated; subjects were given dictionary definitions, and created definitions which were one degree away from the source definitions. These definitions were used in the next round to train subjects using degree one definitions and create degree two definitions, with the final round of this first phase created degree three definitions. This phase generated the training definitions, as well as the scorings for the teachers propagating definitions.

In the testing phase, subjects were presented with a similar test structure, though now with a random word definition from the correct degree level for each of eight tested words. Each subject was tested with two words from each degree level, as well as two untrained words. Scoring was divided in a binary manner, placing results into either the naive or trained set. We intended to link this data back to the score of the definition creator for degrees one through three for analysis; when we returned to the data for analysis we realized that poor data management techniques had eliminated our ability to fully link degree one and two definitions back to their creation source.

Results

We had fifty-four subjects in the final round of the experiment, after the learning phases used to generate definitions. Subjects were aged eighteen to sixty, twenty-eight of the subjects identified as female and twenty-six identified as male. The large majority, forty-nine, were located in the United States, three in India, one in the Philippines, and one in Croatia.

While the intention was to study how ideas were retained and passed through generations of learners, data was not fully retained on the origin of created definitions. Without this data, the

dataset is unable to be modeled as a graph of teachers to learners, and instead looks at correlations of correctness across degrees within subject. We are then testing the null hypothesis that degree distance from source definition does not impact performance after a learning phase. Correlations were studied using an ANOVA, as seen in Figure K; we rejected the null hypothesis due to a P value of less than 0.001. Degree distance for source definition impacts performance after a learning phase, with lower distances correlated with higher performance.

Test fatigue impacted mean percent correct, as shown in Figure I and discussed further in the explanation of figures. Degree 3 definitions post training performed worse than naive state pre training, and naive state after training performed worse than naive state before training. It is possible that the length of our test should be revised in future studies, down from our sixty-four questions.

Mechanical Turk provides an amazing platform for the development of future studies, and an opportunity to propagate those studies programmatically. The toolset for interacting with Mechanical Turk is a realm for improvement, however, providing studies with a high barrier to entry from a technical perspective.

Discussion

We attempted to explore knowledge passing and signal loss in a connected network, but only scratched the surface of potential topics and improvements, both in the realms of experimental design and Mechanical Turk. Better data management techniques and experimental design would result in a more informative and conclusive data set, even without major modifications of our experiment. A better designed data set could be used to explore the signal resiliency of ideas and teachers, rather than only the resiliency of singular definitions.

Before diving into our results, such that they are, it needs to be acknowledged that Mechanical Turk provides a unique and powerful potential platform for further studies which are iterative in design. The customization of Mechanical Turk surveys is limited only by the capabilities of Javascript to feed forward form information. Data gathering could be expanded from our example to include elapsed times, correctness over test progression, more comprehensive demographics, mouse input, question selection ordering, and an array of potential features. In addition to traditional features our experiment could be expanded to preserve a fully connected network, showing the path connecting source and terminating nodes to encompass and preserve the full network, and additionally study further the impact of good and bad teachers on the extended network. Our experiment could also be automated to propagate and present new rounds without manual intervention, deploying the experiment and simply waiting for the complete collection of data while rounds are generated and submitted automatically. In this way a definition's evolution could be studied at a degree depth which would be severely complex in a more traditional model. Mechanical Turk offers unique design potential which eschews traditional experimental models, but suffers from a lack of documentation, understanding, and tools, providing a high barrier to entry and greatly complicating experimental execution.

The area with the largest potential for future exploration and improvement is in data management and experimental design. Our study was limited to providing high level generalizations of how vocabulary understanding decays through succeeding rounds of definition creation, while mainly focused on a single layer of connections. While we look at how a learner conveys ideas relative to their total score and improvement on a specific word and degree, we fail to map out or explore whether these generalization can be expanded across degrees, and how

closely linked strong idea transmission is with the originating individual, regardless of degree. We also fail to dive into the pervasiveness of successfully communicated definitions, and whether specific teachers create definitions that are more or less impactful through successive generations, rather than simply at the first degree.

Another area of potential study lies in exploring whether specific definitions or learners show resiliency beyond the first degree, and exploring signal propagation through the viewpoint of studying successful teachers rather than successful definitions. A fully connected graph would allow study of whether particular teachers communicate ideas with less knowledge decay through succeeding generations, and what properties define these teachers and resilient ideas, allowing us to explore other metrics of the trainee and trainer performance of Ghodsian, Bjork and Benjamin (1997) evaluating the "variety of manipulations that impede performance during training", but "facilitate performance on the long term". There may be factors outside of base scoring and understanding of materials which can be analyzed in assessing the real performance of a teacher. Our experiment only explores the generalities of score through generations, looking at each source not as a sum of answers and successful knowledge transfer, but rather as the data point of a single score, and whether the score trends through successive generations only along one degree of outgoing edges. Our score oriented approach may be additionally informative if tracked as a teacher oriented approach.

Tools like Mechanical Turk provide a great platform for unique experimental design, and, as seen in our experiment, allow a departure from conventional methods of experimental flow and sequencing; this could be leveraged to create experiments which explore novel network effects, better represent individual and independent actors, and which are highly iterative, all

while avoiding the high cost and complexity of a traditional design to model the same. Despite the potential of Mechanical Turk as a research tool, it remains underutilized, poorly documented, and significantly less approachable than more conventional designs.

While our experiment models a graph of communication decay across words, and in overall loss over one degree, we are taking advantage of only a small set of the data which Mechanical Turk can provide. In our case, this was due to limits of understanding regarding the Mechanical Turk API and to a lack of a comprehensive data management strategy from the inception of experiment design. Our research could easily be expanded to examine and answer questions of elapsed time and correctness, demographic effects, and to look at data as a network, not simply as transitions between individual degrees. In hindsight our data representation fails to provide a full understanding of network flow; each edge connecting definitions is considered individually, discarding node information that could be used to link together full paths through the network, overall failing to see the entire path of a definition communicated and connected from source to terminating nodes. To gather this information we would need to add data about each created definition; currently we only consider the word and degrees from source, but this does not provide a full representation of a unique node. We would also need to capture the originating experimental subject for each definition, and the degree of the created definition. With this addition, an experiment and subsequent data analysis could easily follow trends related to multilevel ancestry, with a minimum burden of additional data.

These additions would also allow for a discussion of whether different definitions or originating experimental subjects exhibit differing resiliency to information decay. Our experiment focuses on information passing at only one level, discarding the idea that a specific

teacher or definition might not simply help the following generation to succeed, but may also be able to communicate ideas and concepts in such a way as to be easily transferrable and retainable, allowing and enabling other teachers to better convey core concepts in the following iteration. Further research could explore the work of Delaney, Verkoeijen & Spirgel (2010) that "tests seem to slow forgetting", and the proposal that this may be due to the differences in the strength and routing of created neural pathways during testing against those created during simply training. It may be that the understanding of a definition is not the only factor in successful teaching, and we also have to consider the ability to transfer and articulate in a way that helps others to do the same. The hope would be to explore more fully Wegner's observation that when learning, "each subsequent mistake, as it is corrected, provides the opportunity for transmission of precisely the information the student needs to complete the operation" (Wegner). If learners for a specific task are able to distill and codify information more succinctly and clearly, then this information could see less decay through successive generations. This could be of interest to teaching professions, where the understanding of concepts among teachers is not generally the limiting factor in conveying ideas, but is still used to dictate hiring. It may be that a better understanding of idea conveyance could facilitate new hiring methods, with a focus on the skill of conceptualizing data, rather than outright knowledge of the taught domain.

Tools like Mechanical Turk provide a great platform for unique experimental design, and, as seen in our experiment, allow a departure from conventional methods of experimental flow and sequencing; this could be leveraged to create experiments which explore novel network effects, better represent individual and independent actors, and which are highly iterative, all while avoiding the high cost and complexity of traditional design. Our experiment did not fully

utilize the power of Mechanical Turk. While we explored the idea of propagating information through successive degrees in a graph, it was still explored in serial fashion, generating a set of questions, and using that set to feed the next generated set. Mechanical Turk allows for full API control over an experiment, and can create experiments, gather data, and create new experiments in near real time and with the potential for full automation. Automation was an area where our experiment only scratched the surface of the potential complexity that Mechanical Turk provides. We utilized boto, a Python library simplifying access to Amazon Web Services (AWS). After each round of definition generation completed, we downloaded a CSV file provided by Amazon, which was then formatted and sanitized via Python. We processed this CSV data into Python data structures, using a variety of scripts to extract data into stored pickle and json files before building, uploading, and starting the next round. This script sanitizing and data processing step impacted our final dataset; in our experiment it limited the scope of our data collection methods and impacted our gathered data, though this could be avoided with more careful planning and an API specifically designed around integration with Mechanical Turk. While boto provided an easy interface and low barrier to entry for simple communication with the Mechanical Turk API, it serves primarily as an interface to the whole of AWS, and this development focus is apparent in building experiments. Even a purpose built and Mechanical Turk specific API would not address all the issues of experiment design; in reality the scientific community would greatly benefit from Python and other modules specific to creating, managing, and gathering data from experiments using Mechanical Turk.

Even with improved and automated round generation, our experiment still would not have taken full advantage of the power of Mechanical Turk. The connected APIs allow a much

finer granularity of experimental control, and could do away with the concept of rounds entirely. Rather than thinking about generating all the definitions at degree one, then degree two, etc., it is possible to generate cases in a much more fluid manner, which is also more in line with the underlying graph. Essentially, each completed HIT can be processed directly upon submission, and then used to augment and grow the data set and propagate more HITs. The first survey only has access to degree zero definitions, and therefore can only generate degree one definitions, while the next can now include both degree zero and degree one definitions, and therefore generate definitions for degree one and degree two. With only three total HITs, we could have definitions for all of the tested degrees in our experiment, with successive rounds continuing to expand the tested definition set. This approach would have removed the need for our sixty definition building trials, while generating definitions from subjects with a more varied, comparable, and useful preceding question set.

The problems we faced with regard to Mechanical Turk are also a potential for future work, but on tools for experiments rather than design and execution. While Mechanical Turk provides in theory API access and the potential for complex and automated experiments, in practice the available tools are poorly documented, very general in scope, and clunky in their current implementation when used to create and manage experiments. We very quickly found ourselves outside of documented features, and expending time to build a better interface with Mechanical Turk rather than creating and processing data. A framework for creating Mechanical Turk experiments has the potential to greatly improve experimental design and execution, and would provide a great foundation for future work.

Explanation of Figures

A demo of a sample test from the final round of testing can be found immediately before figure A, and is what is shown in figures A, B and C. Figure A shows a single page of the test, with words to be matched to synonyms. Continuing forward with the test is irreversible, and participants are unable to revise answers. Participants are exposed to only four questions at a time, and tested words within each four question set are unique. Each test was generated randomly, the demo and examples are from one of the randomly generated tests.

Figure B shows our training phase, with a degree one definition at the top, and a degree three at the bottom. You can get an idea of the quality degradation of created definitions here, as well as a potential improvement for future study. If a participant scored poorly on a word overall, their definition would still show up in future tests, potentially propagating early failures into successive definitions. This could be better analyzed from a connected perspective.

Figure C shows the creation of new degree definitions. In the data collection rounds, following the test and prior to collecting demographic information, we collected new definitions from test taking learners. This collection was used for the creation of each of degree one, two and three definitions. A more sophisticated experimental design could have greatly shortened this data creation phase, forward propagating newly created definitions to new tests in real time rather than through rounds. This would have eliminated the need for most of the data collection rounds, which instead could be compressed into as few as three data collection tests building six degree one definitions, four degree two, and two degree three. This would require keeping track of which definition originated from which source definition, and additional controls to expose

each definition equally among tests, as otherwise our data sampling would contain bias towards earlier created definitions.

Figure D shows a selection of words from the Corpus of Contemporary American English, a corpus of 450 million words. These are paired with their part of speech, frequency, and dispersion among the works in the corpus. A word's frequency as well as dispersion are used to calculated its ordering within the word list, as higher frequency word with low dispersion can appear on the list lower than a lower frequency word with high dispersion.

Figure E shows the collection of synonyms for our tested words. Each word is paired with six synonyms, required to avoid repetition of word synonyms for the six total questions per word on each test. Words with fewer than six synonyms were discarded from our experiment, and word synonyms were collected programmatically (Big Huge Thesaurus).

Figure G shows our source definitions, degree zero definitions, taken from Merriam-Webster's Learner's Dictionary (Dictionary API). These definitions were shown during the training phase, and learners exposed to these definitions created degree one definitions.

Figure H shows the disqualification questions used to remove insincere participants who select question answers at random. These screener questions removed fourteen participants in total across data collection and testing rounds.

Figure Ia shows the mean correct answers of varying degree levels, by percent correct. Some of the impacts of test taking fatigue over the course of the test can be seen in this data. Pre-training naive state is indicated as a dotted line, with post-training naive state indicated by the solid line. Pre-training naive state performed better than post-training naive state, even

performing better than post-training degree three learners. Figure Ib shows the same data, marked by mean number correct.

Figure Ja and Jb shows performance across individual words, regardless of training state, both by number and percent correct, showcasing the performance variance of individual words within testing.

Figure K shows the results table of our ANOVA, disproving the null hypothesis that definitions degree distance from source has no impact on testing performance following a training phase. This was disproved with with $P < 0.05$. This analysis could be performed within subject, as all subjects were exposed during testing to two words at each degree level, and two words without any training during a training phase.

Figure L showcases the connected strength of test performance propagation from teachers to learners among available data, while also showing the loss of edge data at early generation rounds. Not all learner created definitions could be reconnected to the originating source and score, as evidenced by missing numbers in the low number of data points at the left. This does show a limited sampling of connection strength between correctness levels at differing degrees, but is not usable for correlational data as this and the following two figures are created from a different number of source data points. Darker lines indicate more results, so a dark line from five on the left to three on the right indicates teachers that answered five questions correctly created definitions of which more tested learners answered three correct. Lighter lines indicate the inverse, that there were fewer connections between teachers of that correctness level and learners of the next correctness level.

Figure M is similar to figure L, but from degree one teachers to degree one learners. This again shows a loss of edge connection data, and so doesn't represent data from all subjects.

Figure N is again similar to figures M and L, but consists of data from all source edges. With a larger data set the correlation of propagation is more visible, showing trends in the strength of idea transmission from degree two teachers to degree three learners.

References

Big Huge Thesaurus. API. https://words.bighugelabs.com/api.php

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. Psychology of learning and motivation, 53, 63-147.

Dictionary API. Merriam-Webster's Learner's Dictionary. http://www.dictionaryapi.com/

Ghodsian, D., Bjork, R. A., & Benjamin, A. S. (1997). Evaluating training during training: bstacles and opportunities. Training for a rapidly changing workplace: Applications of psychological research, 63-88.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In Theories of group behavior (pp. 185-208). Springer New York.

Word Frequency Data: Corpus of Contemporary American English. http://www.wordfrequency.info/

Figure A

Demo vocabulary test available at https://goo.gl/DpGw3B

Figure B

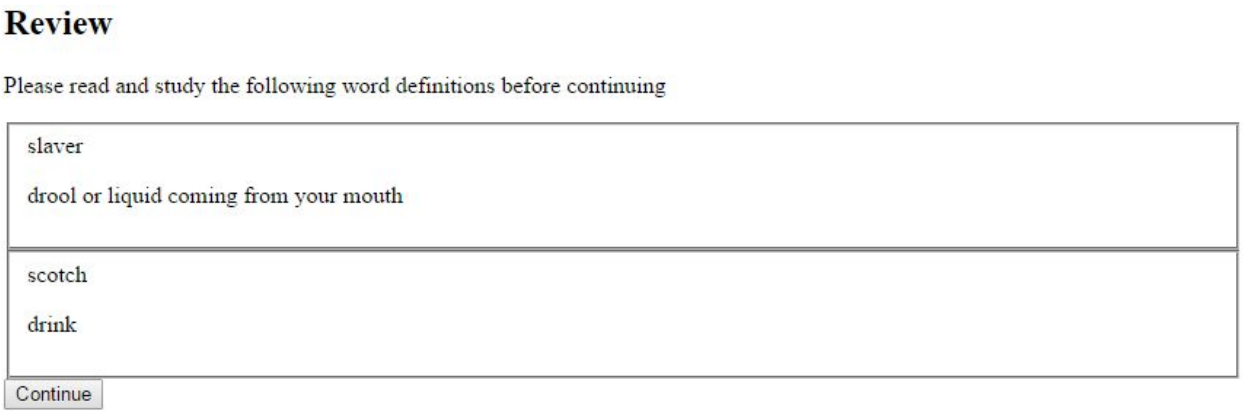
Review
Please read and study the following word definitions before continuing
slaver
drool or liquid coming from your mouth
scotch
drink
Continue

Figure C

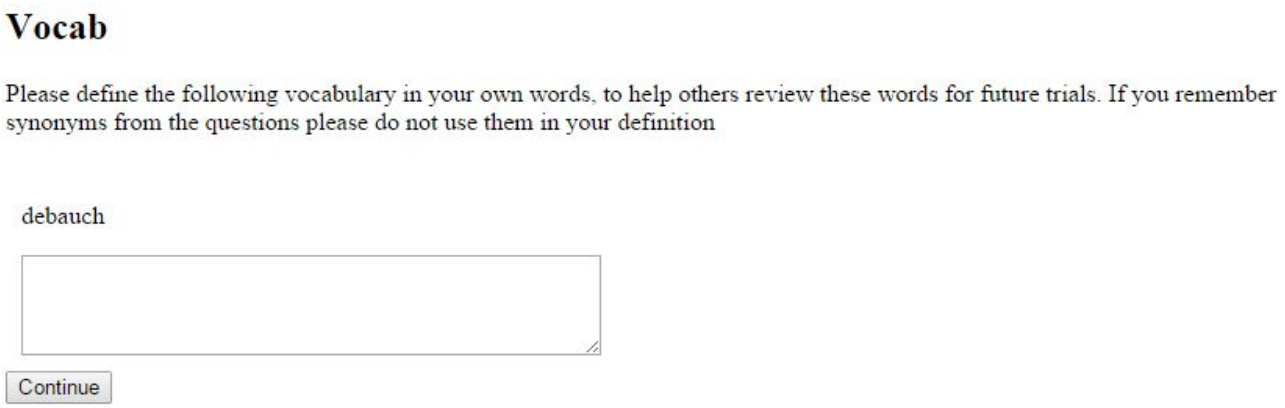
Vocab
Please define the following vocabulary in your own words, to help others review these words for future trials. If you remember synonyms from the questions please do not use them in your definition
debauch
Continue

Figure D

45899	enumerator	n	46	0.70	
45906	mandoline	n	60	0.54	
45913	antinomy	n	43	0.75	
45920	guidestar	n	52	0.62	
45927	shatterproof	j	39	0.83	
45934	bestride	v	39	0.83	
45941	hand-cut	j	39	0.83	
45948	flag-waving	n	40	0.81	
45955	presidentially	r	41	0.79	
45962	female-dominated	j	42	0.77	
45969	plugger n	43	0.75		
45976	rotating	n	39	0.83	
45983	fragrance-free	j	43	0.75	
45990	initiating	j	42	0.77	
45997	eeriness	n	40	0.80	
46004	nonrecognition	n	40	0.80	
46011	two-timing	j	39	0.82	
46018	trustful	j	39	0.82	
46025	ethnicization	n	75	0.43	
46032	lecithin	n	43	0.75	
46039	kahuna n	42	0.76		
46046	unnameable	j	40	0.80	
46053	lumpen j	40	0.80		
46060	tautness	n	39	0.82	
46067	well-accepted	j	39	0.82	
46074	contrastive	j	47	0.68	
46081	fancifully	r	38	0.84	
46088	great-great-grandmother	n	41	0.78	
46095	interleukin	n	48	0.67	

Figure E

picket: [lookout, lookout man, sentinel, sentry, watch, spotter]

befuddle: [confuse, bedevil, confound, discombobulate, inebriate, intoxicate]

besmirch: [defame, slander, smirch, denigrate, smear, sully]

twiddle: [twirl, swirl, whirl, fiddle with, go around, manipulate]

debauch: [corrupt, pervert, subvert, demoralize, debase, profane]

slaver: [drivel, drool, slabber, slobber, dribble, salivate]

scotch: [thwart, queer, spoil, foil, frustrate, baffle]

exfoliate: [break away, break off, cast off, chip, chip off, shed]

shew: [demonstrate, establish, affirm, corroborate, substantiate, support]

carom: [bounce, bound, glance, recoil, reverberate, ricochet]

Figure G

picket: a soldier or a group of soldiers whose duty is to guard something (such as a camp)

befuddle: to muddle or stupefy with or as if with drink

besmirch: to cause harm or damage to (the reputation of someone or something)

twiddle: to turn (something) back and forth slightly

debauch: to lead away from virtue or excellence. to corrupt by intemperance or sensuality

slaver: to allow liquid to drip out of the mouth

scotch: to put an end to (scotched rumors of a military takeover)

exfoliate: to cast off in scales, laminae, or splinters

shew: to give information that proves (something)

carom: a rebounding especially at an angle

Figure H

big: [large,dog,cat,mouse]

small: [little,run,fox,house]

male: [man,philosophy,prime,complex]

easy: [simple,boot,russia,keyboard]

Figure Ia

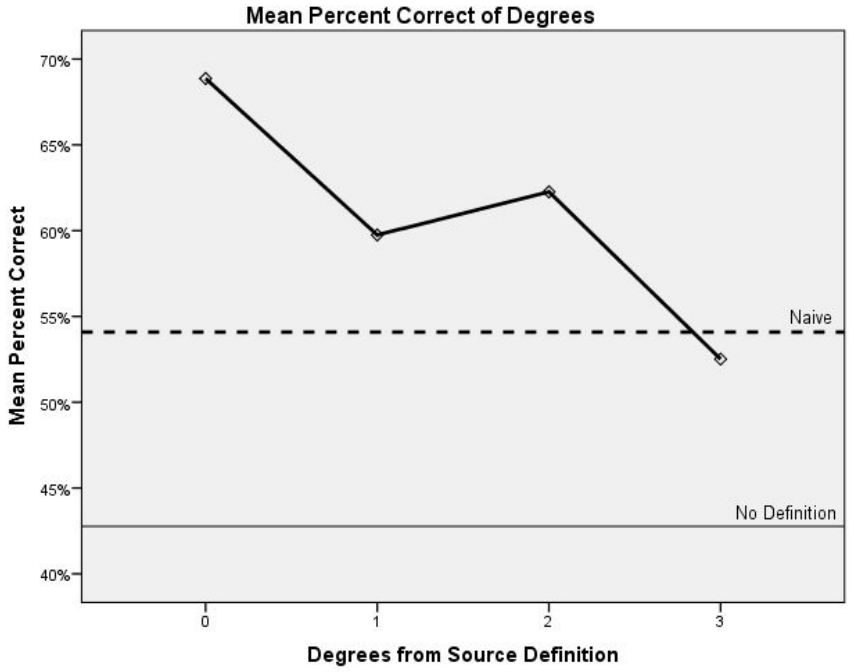

Figure Ib

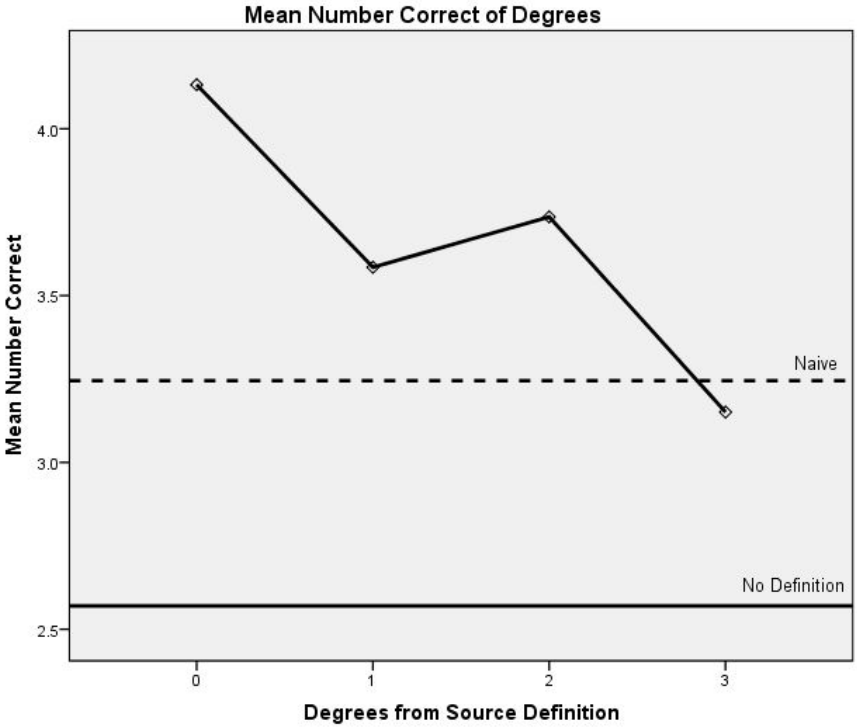

Figure Ja

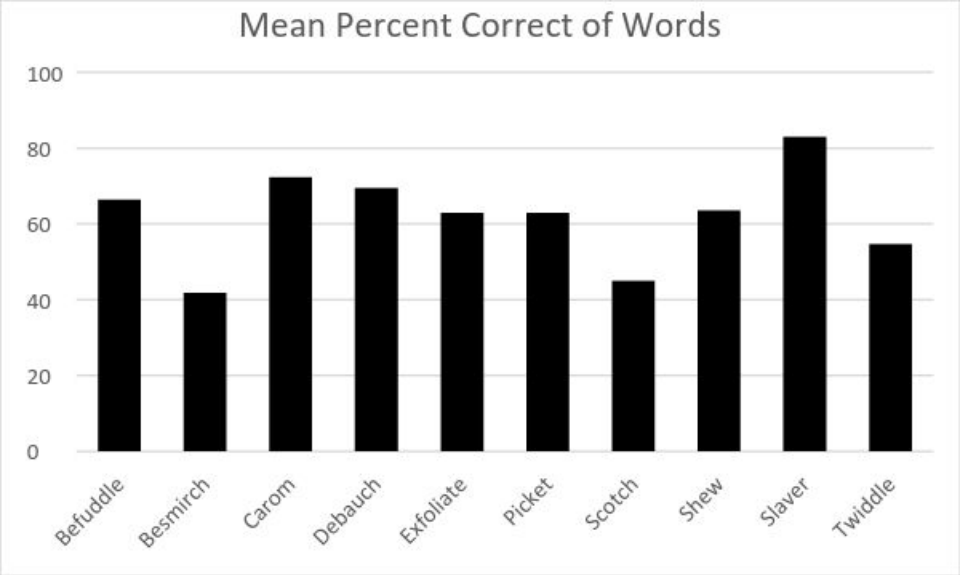

Figure Jb

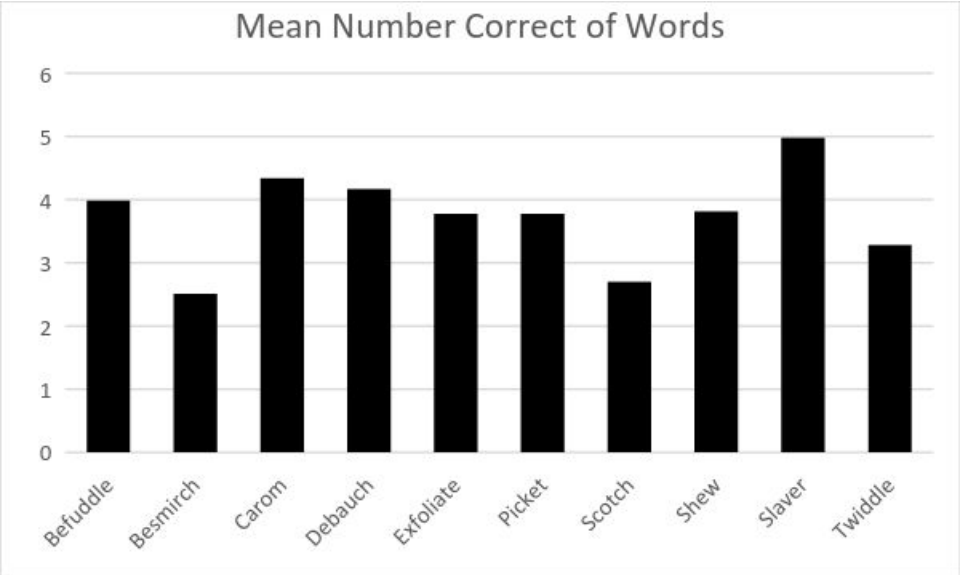

Figure K

Tests of Within-Subjects Effects

Measure: Correct_Answers

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Degree	Sphericity Assumed	73.610	5	14.722	8.690	.000
	Greenhouse-Geisser	73.610	4.220	17.443	8.690	.000
	Huynh-Feldt	73.610	4.631	15.896	8.690	.000
	Lower-bound	73.610	1.000	73.610	8.690	.005
Error(Degree)	Sphericity Assumed	448.923	265	1.694		
	Greenhouse-Geisser	448.923	223.666	2.007		
	Huynh-Feldt	448.923	245.428	1.829		
	Lower-bound	448.923	53.000	8.470		

Figure L

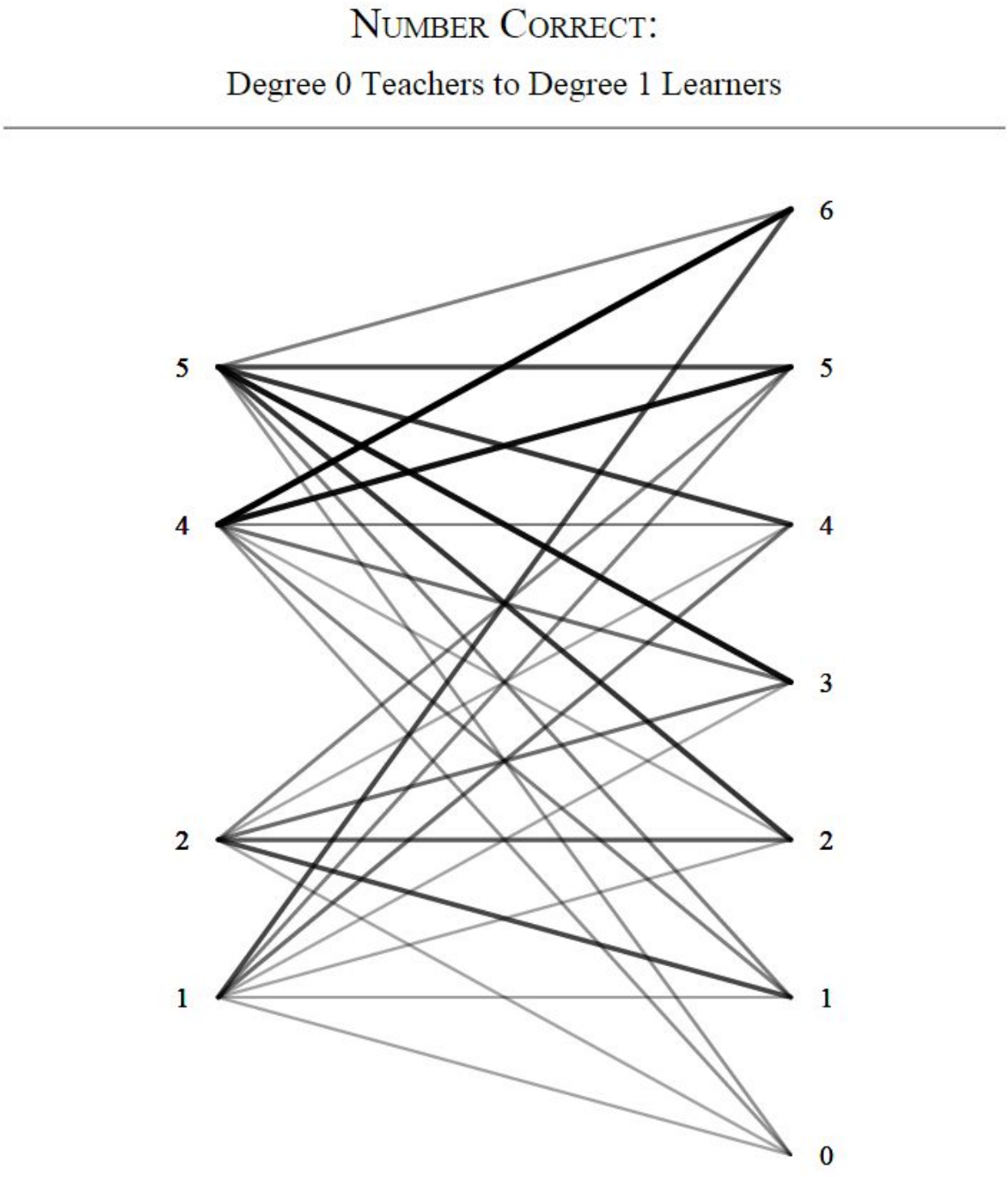
NUMBER CORRECT:
Degree 0 Teachers to Degree 1 Learners
5
4
2
1
6
5
4
3
2
1
0

Figure M

NUMBER CORRECT:

Degree 1 Teachers to Degree 2 Learners

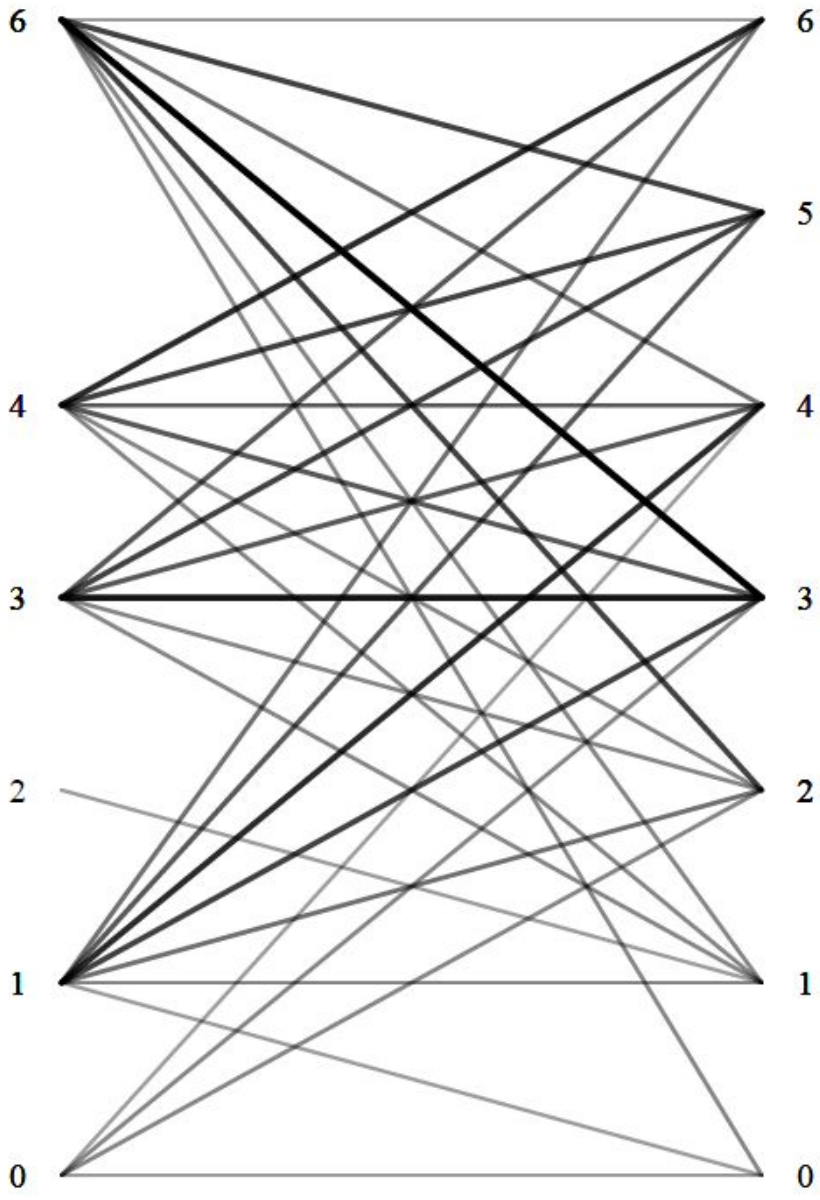

Figure N

NUMBER CORRECT:

Degree 2 Teachers to Degree 3 Learners