北星学園大学文学部北星論集第58巻第１号（通巻第72号）（2020年９月）・抜刷

【研究ノート】

Setting Up Classes for the Magic to Happen

Thomas H. Goetz

研究ノート

Setting Up Classes for the Magic to Happen

Thomas H. Goetz

Contents

[Abstract]

The Foreign Languages Division of Hokusei Gakuen University administered a placement test for all first-year students who wished to continue learning English in their second year. Cengage Learning designed this placement test as a component of the MyELT series. A 51-item assessment was taken by roughly 470 students, evincing a mean of 50% with a more or less normal bell curve. Appropriate language acquisition and advancement occur when students are placed in classes apposite to their current proficiency levels. Clusters of identical scores designate subgroups of students and make drawing distinctions into a subjective or capricious task. The problem of score clustering was sought to be avoided during the sorting of students to level appropriate classes across three faculties. To this end, the weight assigned to questions was reviewed and the uniform, default weight of 1.00 attributed to each question was amended. Questions were assigned differential weightage within an established range to ensure easier ranking decisions. Consequently, score clustering was reduced and became a quick and effortless exercise because enough variation was attained in the test outcomes to simplify the sorting of students. This paper will share the methodology that was undertaken, which yielded favorable and time-saving results that benefited students, teachers, and administrators.

INTRODUCTION

ESL Placement Testing has been around in various systematic forms since the 1940s (Ling, Wolf, Cho, & Wang, 2014). J.D. Brown outlines how placement tests differ from achievement tests is that they are Norm-Referenced as opposed to Criterion-Referenced (Brown, 1989). As its name implies, placement tests are to sort students into level appropriate classes for educational benefit. Not to do so can be detrimental to the student, teacher, and institution (Brown, 1989). While there is no shortage of articles extolling the benefits of placement tests, there is a void when it comes to weighing items according to task ease and difficulty. Are all test items equal? Yet, items appearing within Norm-Referenced tests usually have equal weight. There are two glaring problems that come to mind. When one considers the variety of linguistic and pragmatic tasks on any placement test, is it sound to assume that the tasks presented are on par and equal to each other? If so, then this is the first failure. Placement test administrators need to recognize that the tasks required by the student

Key words：e-learning, placement testing, Moodle, facility index

are different. Should a bottom-up item about the copula have the same weight as a top-down reading/comprehension item? It is argued items of various tasks have their weights modified with respect to the tasks required of the student. And, the second area of concern is the inevitable clustering of scores as a result. Weighing questions to reflect task burden ensures that clustering is largely eliminated, thus taking the guesswork out of line drawing within groups that have the same score. Realizing that ESL placement tests ask learners a variety of questions with a variety of task expectations, it is argued that adapting questions weights to reflect task difficulty benefits the students, instructors, and administrators.

RESEARCH DESIGN

Subjects: The participants include 467 first year students within the faculties of Economics, Social Welfare, and Department of Psychology and Communication. Groups were self-selected based upon departmental membership and that they 1) selected to study English as opposed to other foreign languages, furthermore 2) all wished to continue studying English for their second year. Selection, therefore, was done at the institutional level based upon student preference.

Task and Instrumentation: The placement test used has been provided by Cengage Learning as an adjunct to their MyELT series (citation needed). The students were using the World Link series during their 1st year English Program, a program designed to build foundational English communication skills (Goetz, 2019). The 51-item multiple choice test was administered on Moodle to roughly 467 students, producing a mean score from 50% with a more or less normal bell curve. The test included a variety of Top-Down and Bottom-Up questions that assess grammar, listening, and reading skills, making it an ideal instrument. By default, all items had an equal weight of 1.00 points.

It had been noticed on another Placement Test the problem of score clustering. The sole purpose of any placement test is to sort students into level appropriate classes. With the problem of score clustering, the task of line-drawing for class creation becomes a difficult and problem ridden task.

Brown states that placement tests need to be taken seriously in that initial group assignments can have a major impact on career opportunities and other lifetime related possibilities (Brown, 1989). The question raised is that in light of a variety of linguistic tasks students have to accomplish, why weight all items equally? Within the literature, this issue has not been addressed. What is addressed is the following: Test validity and reliability. The private firm, Professional Testing, Inc, is highly concerned about test validity and seems to take for granted the outcome scores, even within a clustering environment, as normal (Professional Testing, Inc., 2019).

Long et. al. state their placement test of 100 items for a Spanish program is valid. Reported is that it is a great benefit to the students (Long, Shin, Geeslin, & Willis, 2018). Their test was used with over 2,000 students and one can only wonder about score clustering, which went unmentioned. It is

unknown how this problem was resolved.

Size of one's subject pool may have an influence on why score clustering seems elusive in the literature. Ebadi, et. al., consider question types within a placement test environment as a critical factor as for assessing implicit and explicit knowledge among adult learners. With a subject pool of 91 learners, the researcher probably had more than adequate teacher - student familiarity thus rendering score clustering a moot point (Ebadi, Saad, & Abedalaziz, 2014).

The need to address the problem of score clustering is more than apparent leaving one to wonder why the issue is under report if not non-existent. Observances from previous years show that office support staff have taken a medical leave due to work-related stress. The extra time needed to place students with clustered scores has proven to be exasperating. While it is not within the scope to reduce all of the stress related to the beginning of any academic year, one must wonder how and where a favorably different approach can be initiated.

RESEARCH QUESTION

How does a systematic and principled approach for eliminating score clustering look like and is its impact worthy or replication? To answer this, the test items were examined to consider a framework that affords weight credit in balance to item task burden. By weight, it is assumed that weight refers to an attribute of importance or value to an item. Task burden refers to the load associated with the relative ease or difficulty needed to resolve an item. Items were weighted from 1.00 to 1.09 and then, once the data were collected, the test with data duplicated with all weights reset to 1.00 for comparison.

METHOD

Items were identified as either Top-Down or Bottom-Up. Top-Down tasks typically require some sort of inferencing from a graphic, audio text or a reading. Bottom-up tasks are those that are data-driven. Most of the subjects are products of a language learning environment that rewards test takers for achieving high scores on largely Bottom-Up tests.

The framework developed uses both Bottom-Up and Top-Down items with features and distractors, weights, and Facility Index. The Facility Index refers to the (F) or mean score of students on an item. To interpret, the following rubric is a generally accepted guide within the Moodle community. Data are reflected in percentages.

Facility Index (F) Interpretation Table	35% - 65% About right for the average student.
5% or lessExtremely difficult or error.	66% - 80% Fairly easy.
6% - 10% Very Difficult.	81% - 89% Easy.
11% - 20% Difficult.	90% - 94% Very easy.
21% - 34% Moderately difficult.	95% - 100% Extremely easy.

The following tables indicate a range of weights covering a variety of tasks. The Facility Index confirms that, in general, the greater the weight a question was assigned, the more difficult it was for the students. And, the lighter the weight assigned, the easier it was for the students. The Bottom-up items included comprehension, grammar and vocabulary questions with occasional audio text and pictures. The Top-down questions were all text based and included comprehension, pragmatic, and vocabulary questions. Table summaries that follow show example questions with the assigned weights and resulting Facility Index.

Bottom-Up	Item Sample	Media / Task	Distractors	Wt.	Facility Index
Grammar Image	My name ___ Paul. [Picture]	Image of a male.	A. am B. calls C. is D. call	1.01	92% Very Easy
Comprehension Audio	[Audio file] Why is the man worried?	***A:*** *The new teacher seems nice. What do you think?* ***B:*** *Well, I'm worried about some of the topics she wants to cover in the International Relations Class.* ***A:*** *Really? Why?* ***B:*** *They're a lot different than what we studied last term....* ***A:*** *Maybe it won't be so bad.*	A. He likes history more than economics. *B. The topics are different from before.* C. The teacher does not seem nice.	1.07	49% About Right
Pragmatic Text	What do you think of your new roommate?	Select one.	A. She studied at Harvard. B. She has brown hair. C. She is a good cook. D. She works at the bookstore.	1.02	51% Fairly Easy
Grammar Text	A. Who baked the cake? B. It _____ this morning by Alana. It's good, isn't it?	Select one.	A. were baked B. is baked C. is being baked D. was baked	1.03	69% Fairly Easy
Vocabulary Text	I need to _____ $1,000 from my savings account.	Select one.	A. receipt B. check C. withdraw D. deposit	1.05	22% Moderately Difficult
Syntax Text	A: Did you _____? B: No, I didn't. Sorry.	Select one.	A. send a letter at me B. sent the letter to me C. send me the letter D. sent me letter	1.04	51% About Right

TABLE SUMMARY - Bottom-up Items

Weight - Facility Index - Comment	Weight - Facility Index - Comment
1.01 - 92%- Fairly Easy	1.04 - 23% - About Right
1.02 - 51% - Fairly Easy	1.07 - 49% - About Right
1.03 - 69% - About Right	1.05 - 22% - Moderately Difficult

TABLE SUMMARY - The Top-Down Framework:

Top-Down	Item Sample	Media / Task	Distractors	Wt.	Facility Index
Discourse Text	[_____]. Sometimes it's because we like the idea of having a car like the one our neighbor has. Or perhaps we buy clothes and jewelry that we will never wear, but the advertisement at the store made us think about having a different look. Is it so important to have many expensive perfumes at home? If everyone thought for a minute about all the people in the world who don't have enough money to buy food, we would stop buying things we don't need and would start helping those in need.	*Choose the best statement to start this passage.*	A. *I never understood why we spend so much money on things we don't need.* B. We need to help people go shopping more often. C. There is no doubt that shopping is the best solution to all our problems. D. People hate shopping unless it's absolutely necessary.	1.09	30% Moderately Difficult
Comprehension Text	Maurice is a very busy person. He gets up at 6:00 AM and takes a shower. Before going to work, he has breakfast with his friend George. He opens his office at 8:30 AM and helps customers all day. Gaby helps him at the office in the morning. At 7:00 PM he closes the office and goes home. Sometimes he plays golf with his friend Luke on Wednesday evenings.	*Choose the best sentence.* What does Maurice do in the afternoons?	A. He meets with Luke. B. He meets with George. C. He works alone. D. He works with Gaby.	1.07	22% Moderately Difficult
Pragmatic Text	Maurice is a very busy person. He gets up at 6:00 AM and takes a shower. Before going to work, he has breakfast with his friend George. He opens his office at 8:30 AM and helps customers all day. Gaby helps him at the office in the morning. At 7:00 PM he closes the office and goes home. Sometimes he plays golf with his friend Luke on Wednesday evenings.	*Select one.* What does Maurice do in the afternoons?	A. He meets with Luke. B. He meets with George. C. *He works alone.* D. He works with Gaby.	1.04	66% About Right

Inference Text	Our daughter turns 20 this year. She's _____ for her age. Many people like her ideas about the future.	*Choose the best word.*	A. childish B. mature C. modern D. elderly	1.06	29% Moderately Difficult
Grammar Text	Jim is 8 years old. He ____ drive a car.	Select one.	A. can B. *can't* C. knows D. isn't	1.01	83% Easy
Pragmatic Text	Tell someone where you live.	Select one.	a. I work at a bookstore. b. *I live in an apartment.* *c. I am Japanese.* d. China is a big country.	1.04	72% Fairly Easy

TABLE SUMMARY - Top-Down Items

Weight - Facility Index - Comment	Weight - Facility Index - Comment
1.00 - 83% - Easy	1.06 - 29% - Moderately Difficult
1.02 - 72% - Fairly Easy	1.07 - 22% - Moderately Difficult
1.03 - 66% - Fairly Easy	1.08 - 30% - Moderately Difficult

DATA GATHERING

The Placement Test was made available under ideal circumstances within a Moodle environment, for three weeks at the close of the second semester. 467 students took the test. The subjects had up to 30 minutes to finish and in the case that time ran out, data was saved automatically. Once it was completed, data were downloaded to an Excel file. The results were as expected; score clustering was reduced significantly. With the data in hand, a duplication of the Placement Test was made with student data. The weights were then reset to 1.00 each for comparison purposes to see to what extent, if any, differences between the data sets and within the groups would be.

Faculty	Class	Capacity	Data:Set Style	Set CI Total	Set CI Borders	Data:Range Style	Range CI Borders
Econ	F	24	41.18	1	1	41.18	
Econ	F	25	41.18	1	1	41.16	1
Econ	F	26	41.18	1	1	41.16	1
Econ	F	27	41.18	1	1	41.16	1
Econ	F	28	41.18	1	1	41.11	
Econ	F	29	41.18	1	1	41.10	
Econ	F	30	41.18	1	1	41.01	
Econ	F	31	41.18	1	1	40.98	
Econ	F	32	41.18	1	1	40.77	1
Econ	G	1	41.18	1	1	40.77	1
Econ	G	2	41.18	1	1	40.67	
Econ	G	3	41.18	1	1	40.64	

Score Clustering with Set Style vs. Range Style

RESULTS

When the Placement Test items share the same weights, of the 467 participants, 456, or 97.85% shared the same scores. With the ranging weights in place, of the same 467 participants, only 69, or 13.04% shared the same scores. The following is a representative example.

Within the Faculty of Economics, there were 12 students with an identical score of 41.18 and this also formed a cluster on a class border. This score represents a common result from all items having the same weight of 1.00. One has to decide how to place the students; into either the higher or the lower level class.

When item weights cover a weight range according to task, the results are favorably different; less students need consideration. The question remains, however, if by introducing ranging weights, was the data skewed or distorted? In short, no. Initially, two T-Tests were performed with the data set to see if there was a difference between the data with uniform set weights and the ranging weights. The Two Sample t-test showed no significant difference between the means of the groups. In the Matched Paired t-test, there was a difference.

TWO SAMPLE T-TEST - WELCH'S T-TEST

It was observed with the two sample t-test (Welch) test, T distribution, DF=932 (two-tailed), that the average of the Set to 1.00's population is considered to be equal to the average of the Ranging's population. In other words, the difference between the average of the Set to 1.00 and Ranging populations is not big enough to be statistically significant. The Null Hypothesis (H0) is to be accepted given that p-value > α.

P-value

P-value equals 0.874575, (p(x≤t) = 0.562712). This means that if we would reject H0, the chance of type I error (rejecting a correct H0) would be too high: 0.8746 (87.46%). The larger the p-value, the more it supports H0.

The statistics

The test statistic t equals 0.157893, is in the 95% critical value accepted range: [-1.9625 : 1.9625]

x1-x2=0.16, is in the 95% accepted range: [-1.9800 : 1.9800]

Effect size

The observed standardized effect size is small (0.010). That indicates that the magnitude of the difference between the average and average is small. There is no significant difference in this case.

PAIRED T-TEST

A paired sample T-test test was carried out, using a T distribution, DF=466, (two-tailed) to test the Null Hypothesis (H0) that there is no significant difference between the groups in a matched paired environment, or that the items, when weighted with variant and ranging weights, are not different.

It was observed that the average of the ranging group minus the set weight group's population is considered to be not equal to the $\mu 0$. In other words, the difference between the averages of the two is big enough to be statistically significant.

P-value

The p-value equals 0.00694017, ($p(x \leq t) = 0.00347009$). This means that the chance of type1 error (rejecting a correct H0) is small: 0.006940 (0.69%). The smaller the p-value the more it supports H1.

The statistics

The test statistic t equals -2.711756, is not in the 95% critical value accepted range: [-1.9651 : 1.9651]

x=-0.16, is not in the 95% accepted range: [-0.1200 : 0.1200]

Effect size

The observed standardized effect size is small (0.13). That indicates that the magnitude of the difference between the average and $\mu 0$ is small.

ANALYSIS OF VARIANCE

Consideration of a means comparison reassures that the manipulation of the weights did not change the means differences at the population level. Within and between departments is another consideration. To address this, an Analysis of Variance was carried out within the four groups and between the two weight treatments, 1) where the weights were set evenly (Even) and 2) where the weights varied over a range (Varied). The profile of the subjects appears below.

SUBJECTS

Weights set to 1.00	Econ 1.00: 161 values	SW 1.00: 146 values	Com 1.00: 61 values
Ranging Weights	Econ Ranging: 161 values	SW Ranging: 146 values	Com Ranging: 61 values

The One Way ANOVA test, using F distribution df(7,1038) (right-tailed) was used to test against the Null Hypothesis that there is no significant variation between the groups where p-value$< \alpha$, H0 is rejected. It was observed that some of the groups' averages are to be considered as not equal. In other words, the difference between the averages of some groups is big enough to be statistically significant.

F table

Source	Degrees of Freedom	Sum of Squares	Mean Square	F statistic	p-value
Groups (between groups)	5	11046.129204	2209.225841	9.763503	4.18033e-9
Error (within groups)	928	209982.171670	226.273892		
Total	933	221028.300873	237		

P-value

With the p-value equal to 4.18033e-9, [p(x ≤ f) = 1.00000], the chance of a type I error (rejecting a correct H0) is small: 4.180e-9 (4.2e-7%). With the smaller the p-value the stronger it supports H1, that there is a meaningful connection.

The statistics

The test statistic f equals 9.763503, is not in the 95% critical value accepted range: [-∞ : 2.2237]. When seen in the aggregate, there is not enough variation between and within groups to say that there is a significant difference in all cases.

FINDINGS

It was found that this new approach of reweighting questions according to task burden was well received by the office staff members and faculty. In fact, many had no idea what was going on, just they worked with a data set that was easy to manage. Within a couple of hours, class creation was completed leaving time for other equally important and timely jobs. While there were some observable differences, they were not enough to dismiss this initiative as an endeavor inherently unfair to the students. With favorable results it would be interesting to see how re-weighing previous years'

Placement Tests would do.

CONCLUSION

ESL Placement Testing has been around for a long time and such tests are to be taken seriously. Placement tests need to be seen in a wider picture, one that not only includes students, but teachers and administrators as well. Re-weighing a placement test to reduce score clustering is a reasonable endeavor. The benefits outweigh the demerits. The benefits include a reduction of time needed to create level appropriate classes and greater accountability within the process of class creation. Items of various tasks should have their weights modified with respect to the tasks required of the test taker. This greatly reduces score clustering. With clustering largely eliminated, much of the guesswork with line drawing in problematic areas is also dramatically reduced. The benefits the students is that they will be placed in level appropriate classes, instructors can expect uniform groups, and administrators can accomplish class creation faster than before with less work related stress.

ACKNOWLEDGEMENTS

This research was funded by specifically designated research funds from Hokusei Gakuen University. It is with sincere thanks that more research may be undertaken in a timely manner.

REFERENCES

Brown, J. D. (1989). Improving ESL Placement Tests Using Two Perspectives. TESOL Quarterly, 23(1), 65–83.

Ebadi, M. R., Saad, M. R. M., & Abedalaziz, N. (2014). Explicit Form Focus Instruction: The Effects on Implicit and Explicit Knowledge of ESL Learners. Malaysian Online Journal of Educational Sciences, 2(4), 25–34.

Goetz, T., Allison, J., Nishihara, A. (2019). 2019 English I-II Student Guide for World Link Book 1.

Ling, G., Wolf, M. K., Cho, Y., & Wang, Y. (2014). English-as-a-Second-Language Programs for Matriculated Students in the United States: An Exploratory Survey and Some Issues. Research Report. ETS RR-14-11. ETS Research Report Series.

Long, A. Y., Shin, S.-Y., Geeslin, K., & Willis, E. W. (2018). Does the Test Work? Evaluating a Web-Based Language Placement Test. Language Learning & Technology, 22(1), 137–156.

NGL Placement Test, (2016). Retrieved July 19, 2019 https://myelt.heinle.com/ National Geographic Learning.

One Way ANOVA. (2019). Retrieved May 10, 2019, from http://www.statskingdom.com/180Anova1way.html

Paired T-Test calculator. (2019). Retrieved July 16, 2019, from http://www.statskingdom.com/160MeanT-2pair.html

Professional Testing, Inc. (2019). How do you Determine if a Test has Validity, Reliability, Fairness, and Legal Defensibility? Retrieved from http://www.proftesting.com/test_topics/pdfs/test_quality.pdf

Two Sample T-Test (Welch's T-test). (2019). Retrieved July 16, 2019, from http://www.statskingdom.com/150MeanT2uneq.html