e-TEALS no. 5 | 2014

An e-journal of Teacher Education and Applied Language Studies

Method-effect on Test-takers' Performance and Confidence in Language Tests

Anna L. MOUTI and George S. YPSILANDIS
| Aristotle University of Thessaloniki

Abstract | In a language test, the number of correct answers is commonly used as the only source of information that defines the result. The question raised is the following: do all task types accommodate all test-takers in the same way? This paper attempts to investigate this relationship by relating correctness and confidence in the answers provided with the task typology. Two achievement tests in two versions (same language content – different task type) were used while test-takers were asked to mark their degree of confidence for each item on both versions. Data obtained through this research were: 1) number of correct answers, and 2) degree of confidence in different task types.

Through the correlations of the variables examined (task typology, correctness and confidence) several findings were registered concerning the relationship between accuracy and confidence in different task types. The task type typology may be seen as a moderating factor which could be responsible for possible variations in accuracy and confidence scores.

Key words | language test performance, confidence, method effect, test question formats, task types

Citation: Anna L. Mouti and George S. Ypsilandis, "Method-effect on Test-takers' Performance and Confidence in Language Tests." *e-TEALS: An e-journal of Teacher Education and Applied Language Studies* 5 (2014): 65-81. ISSN 1647-712X

1. Introduction

The driving force behind this research is the need for a more precise and fair language measurement. When it comes to a language test, it seems that the number of correct answers is the only source of information that defines the result. This approach has only recently been challenged. In Schraw's words: "Effective Test-taking depends on two important skills: selecting correct responses to test questions and monitoring one's performance accurately" ("The Effect of Generalized Metacognitive Knowledge" 135). It may be possible to suggest that the process and measurement of self-monitoring is enhanced through the measurement of confidence. Indeed, Stankov and Lee consider confidence as an individual difference which could be defined as "a systematic tendency that leads one to act in a particular way because it reflects a belief, a faith in oneself" (962). Registering confidence may be another variable (adding to the test-taker profile) which affects test performance by raising metacognitive strategy awareness but also by providing metacognitive information for a more fair and precise scoring.

In the field of Applied Linguistics it was Yule who first traced the notion and the need for the study of confidence. Yule, Yanz and Tsuda "set out to quantify, in a limited way, the effects of that one important affective variable called confidence in one area of the language learner's performance" (475). Their experiment involved the following procedure: each time the participants chose an answer, they had to indicate how confident they were about the correctness of their answer. This approach is adopted in the framework of this study setting language test performance and confidence as the two dependent variables during the test-taking process.

In the discussion for a more precise and fair language testing procedure Bachman claimed that "a major concern in the design and development of language tests is to minimize the effects of the factors that are not part of the language ability" (*Fundamental Considerations* 166). Further, Bachman and Palmer stated that "[a] number of language testers have indicated that we should attempt to design our tests to elicit test-takers' best performance. We believe that one way to do this is to design the characteristics of the test task so as to promote feelings of comfort or safety in test-takers that will in turn facilitate the flexibility of response on their part" (66). In order to

guarantee a precise and fair measurement of the two variables stated above, throughout the study but also in real-life situations and authentic test-taking settings, it was attempted to examine another key independent variable: the question format. Possible positive correlations between the "question format" and language test performance and confidence could place under consideration fairness of language test scores but also the degree to which task-types "accommodate" the test-takers.

2. Question Formats

The independent variable of our research is the task typology and question format used in language tests. More specifically the study concentrates on closed-type tasks, where a fixed response is required. Some of the most frequent closed-ended task types used in language tests are the following:

Closed-ended Test Question Formats	
Selected-response	True/False
	Multiple-Choice
	Matching
Constructed-response	Gap-Filling
	Short answer

Table 1. *Close-ended Test Question Formats*

Douglas sees a test as "a measuring device, no different in principle from a ruler, a weighing scale or a thermometer" (2). For Douglas, "a language test is an instrument for measuring language ability" although the author expresses the following doubt: "In what sense can we actually measure a concept as abstract as *language ability*?" (2). Although it would require multiple papers to provide an answer to this question, it is customary to measure language ability though a number of close-ended questions which are accepted to be used to elicit test-takers' responses. Any estimate of language ability is based on the provided answers; to locate the degree of knowledge in a rather conventional and to a certain extent arbitrary manner. As presented in the

above table these responses can be either selected or constructed. Another term used for constructed-response is provided by Purpura, who makes the distinction between selected-response tasks and limited production tasks which "elicit the response embodying a limited amount of language production" (134). Our research hypotheses will rotate around multiple-choice questions and gap-filling ones as these two are used as representative task types of the selected and constructed response tasks.

Multiple-choice (MC) questions are the most commonly used to measure language competence because they are quick, economical and straightforward to score. A typical MC test item consists of two basic parts: the stem (a question or a problem to be solved) and a list of possible answers, which usually contains one correct (or "best") answer and a number of incorrect options (distractors). Burton et al. give six types of MC items (12-15):

(a) Items of the *single-correct-answer* variety

(b) Items of the *best-answer* variety

(c) Items of the *negative* variety

(d) Items of *multiple-response* variety

(e) Items of the *combined-response* variety, and

(f) Items of *multiple true or false* variety.

Some of the constructed response questions are gap-filling items which are also frequently used in testing. There is a variety of these questions like the cued (or guided) gap-filling questions. In these "the gaps are preceded by one or more lexical items or cues which must be transformed in order to fill the gap correctly" as indicated by Purpura (136). Purpura further suggests another type of gap-filling task: the cloze. In this task type every fifth, sixth or seventh word is deleted and the test-takers are asked to fill in the gap. It was Wilson Taylor in 1953 who first introduced the cloze procedure but there have been a number of variations since then: the standard cloze test, the modified cloze test, the multiple-choice cloze test, the c-test and the cloze elide. The modified

cloze test might be otherwise expressed as selective-deletion gap-filling, setting the "selective" as against random (Weir, Vidakovic, and Galaczi 158).

3. "Method-effect" on Performance and Confidence

Task types have been thoroughly examined in the past and there is a lot of research concerning the "method effect" on language test performance. Alderson, Clapham and Wall offer a working definition of the method effect in that "the method used for testing language ability may itself affect the student's score" and further claim that "its influence should be reduced as much as possible" (44).

Bachman initially discussed a framework for the facets of test methods that may affect language test performance (*Fundamental Considerations* 119). These facets were grouped into five sets: the facets of the testing environment, the facets of the test rubric, the facets of the input the test taker receives, the facets of the expected response, and the relationship between input and response. This was further developed in Bachman and Palmer (47-57) in a framework of language test task characteristics (setting, rubrics, input, expected response and relationship between input and response) where a suggestion is made in that test creators are expected to "design the characteristics of the test task so as to promote feelings of comfort or safety in test-takers that will in turn facilitate the flexibility of response on their part" (66). Shohamy and Wolf found that test methods influenced language test performance and that multiple-choice questions were easier than open-ended questions. This claim was investigated by several researchers. Tsagari investigated the above claim empirically and compared the effects of two test formats (free response and multiple choice) on two reading comprehension tests with identical content. Their findings revealed that method effects can be a source of variation of scores and may even measure different abilities. Cheng in research on listening comprehension came to a similar conclusion in that multiple-choice cloze was easier than regular (distinct) multiple choice and that the open-ended questions were found to be more difficult. In the research conducted by Zheng, Cheng and Klinger, participants achieved a higher percentage of correct responses in multiple-

choice questions and lower in constructed-response questions in reading comprehension. Liu, in 1998, used multiple-choice questions, true or false questions and short answer questions and the results showed that there were significant differences among the scores elicited by the three different test methods with short answer questions being the most difficult (147). Later, Liu conducted a similar research with reading comprehension and concluded that different test methods affect students' scores in this receptive skill. In this research, the gap-filling test was considered the most difficult, while multiple-choice questions and short answer questions were easier. In'nami and Koizumi conducted a meta-analysis on the effects of multiple-choice and open-ended formats on L1 reading, L2 reading, and L2 listening test performance. Fifty-six data sources were located and the results indicated that multiple-choice formats are easier than open-ended formats in L1 reading and L2 listening, with the degree of format effect ranging from small to large in L1 reading and medium to large in L2 listening. Currie and Chiramanee investigated the effect of multiple-choice format, compared with a constructed-response test on equivalent language test items. The scores of the two tests were found to be highly correlated but only 26% of the answers (either correct or wrong) were the same, with the multiple-choice questions being easier. In a recent research conducted by M. Salehi and H. Bagheri Sanjareh, response format was also found to be responsible for variations in language test performance. In general, researchers have proved that in cases that question format is found to be responsible for score variations, selected response type is considered easier than the constructed response type.

Although the impact of question format on language test performance was investigated by a great number of researchers in the last twenty-five years, the literature review about the impact of question/test format/method on language test confidence came up with very limited results. Findings of Pallier et al. show that "when one is answering a question, one's level of confidence is sensitive to question format" (262). Pallier et al. were based on previous research conducted by Kohler, who reached the conclusion that questions which ask for a hypothesis to be expressed generate less confidence than those which provide some alternatives for evaluation. Of course Kohler's research was not implemented in a language test setting but was part of a psychology

experiment. However, it may be possible to conclude that hypothesis generation is nearer to open-ended, cloze or short-answer questions while the alternatives could be considered to be linked to multiple-choice. Thus, Kohler's results could be implemented in a language testing setting. As is implied by the limited number of sources on this issue, further research would need to be conducted.

4. Research Method

4.1. Hypotheses-Design

In this study there are three variables under investigation: Test Question format as an independent variable in relation to two dependent variables, language test performance and language test confidence. More specifically, two research hypotheses were tested:

1. There is a relationship between task-types/test question formats and language test performance.
2. There is a relationship between task-types/test question formats and language test confidence.

In order to investigate the above two hypotheses a quasi-experimental realistic research was designed. The subjects were asked to complete two sets of test items, with the same language content and different task type/question formats. In addition the participants had to mark their level of confidence on a 100mm bar, indicating how confident they were feeling about the correctness of their answer on each item.

4.2. Instruments

The instruments used for measuring language test performance were 6 sets of language test items of constructed response and 6 sets of language test items of selected response. The sets of items were used to test linguistic knowledge relevant to grammar and vocabulary and were part of a longer test used as the final exam of two EAP language courses. There were 6 sets of

items – presented in the first stage of the research – in a constructed response format. More specifically there were 5 cued gap-filling tasks and 1 gap-filling task (selective deletion gap-filling/modified cloze), checking meaningful grammatical and lexical items. During the second phase of the research, the subjects were asked to complete the same sets of items in a different test format, i.e. multiple choice. Based on the studies by Rodriguez and In'nami and Koizumi, test items were constructed with stem equivalency. This was used to examine the impact of task typology on performance (format effects). A similar study using a stem equivalent format was used by Currie and Chiramanee. All the test items were statistically analysed and the facility index was between acceptable borderlines, between 0.48 and 0.82. The test items were scored dichotomously, both the multiple-choice and the gap-filling ones by "scoring for the contextual appropriateness i.e. to count as correct any word that fully fits the total surrounding both syntactically and semantically (Weir, Vidakovic, and Galaczi 70).

The process of confidence marking was implemented in 4 sets of items of constructed response and 4 sets of language test items of selected response. For the confidence measurement, a 100mm confidence bar was adopted for each item on which the subjects could mark "*how confident did they feel about the correctness of their answer*". The subjects were asked to mark their confidence using the bar (ravdos) suggested by P. Kambaki-Vougioukli and T. Vougiouklis. The adoption of the Vougioukli and Vougiouklis bar (ravdos) is claimed to be simpler and easier for both the subjects of an empirical research study and the researcher. Its advantages lie in both the design stage and the processing of the results as it is more flexible than the typical Likert scale.[1] In Kambaki-Vougioukli et al. (82) and Vougiouklis (20) it is clearly stated that:

> In every question, substitute the Likert scale with "the bar" whose poles are defined with "0" on the left and "1" on the right:
>
> 0_____1
>
> The subjects/participants are asked, instead of deciding and checking a specific grade on the scale, to cut the bar at any point they feel best expresses their answer to the specific question.

4.3. Participants

All the participants were first and second year students, of Greek nationality and Greek mother tongue, aged 18-20. Ninety-eight of those were male and 102 female. The majority hold a B2 certificate and shared common educational characteristics, concerning the same educational background and similar EAP experience in their university department. They were all familiarized with language testing procedures and they have all attended English language courses during their primary, secondary and tertiary education.

Four groups of a total of 200 participants (separated according to the EAP course they were attending and who were assessed)[2] completed 12 sets of items in total, 6 of constructed-response and 6 of selected-response. The first group (55 students) completed 2 sets of items, the second group (48 students) 6 sets of items, the third group (65 students) completed 2 sets of items and the fourth group (32 students) completed 2 sets of items during the two stages of the research.

The first two groups (a total of 103 students) completed the 8 sets of items and they also marked their degree of confidence on the 100mm bar after each response they gave on the test.

5. Results

The research product was two series of items, completed and marked in terms of the degree of confidence for each item. Thus two sets of scores – the performance/accuracy scores (degree of accuracy) and the confidence score (degree of confidence) – were obtained and all of them were converted into percentages to create homogeneity in the dataset.[3] These two series of scores were further divided into four sets of scores. Two sets of scores (accuracy and confidence) for the constructed-response items and two sets for the selected-response items of the 8 sets of items examining both accuracy and confidence are presented in the following table. As expected, MC presents, overall, higher performance and confidence scores/percentages than the constructed-response, leading to a first conclusion that task types/test item format is responsible for variations on accuracy and confidence scores.

	Constructed-response (Mean)	**Selected-response** (Mean)
Performance	62.45%	74.49%
Confidence	65%	72 %

Table 2. *Performance & Confidence – Mean*

5.1. Test Question format * Language Test Performance

In the following table we can see the descriptive statistics of all the tasks used in the research procedure. It is obvious that all the multiple-choice tasks may be considered easier than the gap-filling ones as they provide a higher score. Although it was not included in our research hypotheses to examine the variations of gap-filling exercises in relation to language test performance, we can assume that there is no impact of the gap-filling variation on performance (cued gap-filling and selective gap-filling).

PERFORMANCE		N	Mean	Std. Deviation
P1	MC[4]	55	79.04	18.965
P2	CGF	53	56.98	25.178
P3	MC	55	77.53	18.989
P4	CGF	55	67.25	21.696
P5	MC	55	79.05	18.154
P6	SGF	55	67.93	17.513
P7	MC	48	81.88	19.388
P8	CGF	48	57.62	27.919
P9	MC	65	60.80	18.924
P10	CGF	72	55.00	21.189
P11	MC	32	74.38	28.391
P12	CGF	32	47.50	32.429

Table 3. *Language Test Performance Scores – Descriptive Statistics*

The scores were further analysed in each pair of the sets of items examined. As shown in the following table there is a correlation between all pairs showing high reliability between the two

formats, which implies that the test-items are measuring the same construct. It should be noted that the correlations (although they are considered statistically significant) do not reach very strong associations (0.679 at the highest). Further, strong differences in the scores obtained were registered, showing that the differences in scores would be very important when criterion-referenced scores are concerned. That means that the students' ranking (norm-referenced situations) would be the same but when it comes to a decision taken based on the total score according to a pass/fail cut-off point a strong variation would occur. The selected-response items were easier for the participants and thus the facility index for the selected-response items was higher than the constructed-response one, resulting in a higher total score for the multiple-choice test items.

	Mean	Paired Difference	Correlation	Sig.
P1-P2 (MC-CGF)	79.51-56.98	22.528 (.000)	0.569	0.000
P3-P4 (MC-CGF)	77.53-67.25	10.273 (.000)	0.679	0.000
P5-P6 (MC-SGF)	79.05-67.93	11.127 (.000)	0.410	0.002
P7-P8 (MC-CGF)	81.88-57.63	24.250 (.000)	0.546	0.000
P9-P10 (MC-CGF)	60.80-55.00	5.015 (.056)	0.430	0.000
P11-P12 (MC-CGF)	74.38-47.50	26.875 (.000)	0.678	0.000

Table 4. *Question Format * Performance Scores*

Similar correlations were reached in Danili and Read, Johnson and Ambusaidi, and Currie and Chiramanee for selected and constructed response items. Additionally, MC questions were found to be easier and thus resulting in a higher score than the Constructed-response questions supporting evidence of previous studies mentioned in the Literature Review.

5.2. Test Question format * Language Test Confidence

Confidence scores were also analysed in detail, but the results were not similar to the performance scores. It is evident from the descriptive statistics that in all the cases multiple-choice degree of confidence was higher than in gap-filling. However, the differences were not substantial.

CONFIDENCE		Mean	N	Std. Deviation
C1	MC	78.10	49	20.997
C2	CGF	72.90	49	16.989
C3	MC	75.26	47	19.332
C4	CGF	73.43	47	26.225
C5	MC	73.09	46	19.723
C6	SGF	67.67	46	20.978
C7	MC	79.18	39	21.541
C8	CGF	62.64	39	24.508

Table 5. *Language Test Confidence Scores – Descriptive Statistics*

As shown in the following table there is a strong correlation between all pairs of questions showing high consistency between the two formats of questions. In addition, there are differences between the confidence scores obtained by the selected and the constructed response items, showing that the question format could not be considered as an important factor responsible for variations on confidence scores. These findings lead us to assume that factors other than the question format could be responsible for variations in confidence scores. The facility index, the performance/accuracy scores and the knowledge of the language items/content could be one of them but also some inherent characteristic of the test-taker. Therefore it was attempted to calculate the correlation between different pairs. A number of correlations were found to be very strong showing that there might be some personal trait related to self-confidence.

	Mean	Paired Differences	Correlation	Sig.
C1-C2 (MC-CGF)	78.10-72.90	5.204 (0.053)	0.552	0.000
C3-C4 (MC-CGF)	75.26-73.43	1.830 (0.573)	0.566	0.000
C5-C6 (MC-SGF)	73.09-67.67	5.413 (0.118)	0.361	0.014
C9-C10 (MC-CGF)	79.18-62.64	16.538 (0.000)	0.769	0.000

Table 6. *Question Format * Confidence Scores*

Similar findings were recorded in Pallier et al.: MC test items gathered higher confidence scores although in our research the differences were not statistically significant in all the cases.

6. Conclusions

Our initial hypotheses have only been partially supported by the evidence. In particular, our findings revealed that task-type/question format could be a source of variation of scores while it could not be considered as a determining factor responsible for variations in confidence scores. It should be emphasized that there was a strong correlation between the scores obtained by the selected and constructed-response question formats, showing that there is no significant impact of question formats on the ranking of students. As indicated by Farhady,

> in a norm-referenced situation, the increase at the level of the scores would not influence the ranking of the students, i.e. all testees will score higher on the multiple-choice form than they will on the open-ended form. In a criterion-referenced situation however, where there exists a predetermined criterion for the students to meet, low scores would hurt those at the borderline. (222)

In norm-referenced situations the increase at the level of scores, which could be attributed to question format, may not have a significant impact while in criterion-referenced situations the choice of the question formats would need to be carefully examined and implemented. Alternatively, any possible variations would need to be supported by equivalent measurements.

The non-significant amount of confidence variation which could be attributed to question formats implies that participants feel equally comfortable with the question formats under investigation, expressing some kind of extra confidence when dealing with multiple-choice questions, something which could be attributed to the higher facility index.

It should be mentioned that the present study may be subject to some limitations as there were different groups of participants responding to different test-items and not one single group or different groups responding to the same set of items. This was a result of the actual setting and a conscious decision not to jeopardize authenticity of the conditions under which the research was conducted. In support of this decision, Cormick suggested that "it is still unclear whether laboratory-based findings from metacognitive studies can be generalized to actual classroom tasks" (qtd. in Nietfeld, Cao, and Osborne 11). Most certainly this research design could be considered as an implementation of a small-scale replication study which is offered to add to the discussion on the topic.

Notes

[1] The issue of the bar has been investigated in a number of papers (Kambaki-Vougioukli and Vougiouklis, Kambaki-Vougioukli et al. and Vougiouklis).

[2] It should be mentioned that the variables EAP language course and level of proficiency were not proved to affect their overall language test performance.

[3] As mentioned also by Bachman (*Statistical Analyses* 39) for the criterion-referenced language test scores. A similar procedure was also followed for the confidence scores by Stankov (128) and Schraw et al. (434).

[4] MC is an abbreviation for Multiple-Choice, CGF for Cued Gap-Filling, and SCF for Selective Gap Filling.

Works Cited

Alderson, Charles J., Caroline Clapham, and Dianne Wall. *Language Test Construction and Evaluation.* Cambridge: Cambridge UP, 1995.

Bachman, Lyle F. *Fundamental Considerations in Language Testing*. Oxford: Oxford UP, 1990.

- - - . *Statistical Analyses for Language Assessment*. Cambridge: Cambridge UP, 2004.

Bachman, Lyle F., and Adrian S. Palmer. *Language Testing in Practice*. Oxford: Oxford UP, 1996.

Burton, Steven J., Richard R. Sudweeks, Paul F. Merrill, and Bud Wood. *How to Prepare Better Multiple-choice Test Items: Guidelines for University Faculty*. Brigham Young University Testing Services and Department of Instructional Science, 1991. Web. 12 March 2014 <http://testing.byu.edu/info/ handbooks/betteritems.pdf>

Cheng, Hsiao-fang. "Comparison of Multiple-Choice and Open-Ended Response Formats for the Assessment of Listening Proficiency in English." *Foreign Language Annals* 37.4 (2004): 544-53.

Currie, Michael, and Thanyapa Chiramanee. "The Effect of the Multiple-choice Item Format on the Measurement of Knowledge of Language Structure." *Language Testing* 27.4 (2010): 471-91.

Danili, Eleni, and Reid Norman. "Assessment Formats: Do They Make a Difference?" *Chemistry Education Research and Practice* 7.2 (2005): 64-83.

Douglas, Dan. *Understanding Language Testing*. New York: Routledge, 2014.

Farhady, Hoossein. "Varieties of Cloze Procedure in EFL Education." *Roshd Foreign Language Teaching Journal* 12 (1996): 217-29.

In'nami, Yo, and Rie Koizumi. "A Meta-analysis of Test Format Effects on Reading and Listening Test Performance: Focus on Multiple-choice and Open-ended Formats." *Language Testing* 26 (2009): 219-44.

Johnstone, Alex. H., and Abdullah Ambusaidi. "Fixed Response: What Are We Testing?" *Chemistry Education: Research and Practice in Europe* 1.3 (2000): 323-28.

Kambaki-Vougioukli, Penelope, and Thomas Vougiouklis. "Bar Instead of Scale." *Ratio Sociologica* 3 (2008): 49-56.

Kambaki-Vougioukli, Penelope, Alexander Karakos, Nikolaos Lygeros, and Thomas Vougiouklis. "Fuzzy Instead of Discrete." *Annals of Fuzzy Mathematics and Informatics* 2.1 (2011): 81-89.

Koehler, Derek. "Hypothesis Generation and Confidence in Judgment." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (1994): 461-9.

Liu, Feng. "The Effect of Three Test Methods on Reading Comprehension: An Experiment." *Asian Social Science* 5.6 (2009): 147-53.

Nietfeld, John L., Li Cao, and Jason W. Osborne. "Metacognitive Monitoring Accuracy and Student Performance in the Classroom." *Journal of Experimental Education* 74.1 (2005): 7-28.

Pallier, Gerry, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard Roberts. "Individual Differences in the Realism of Confidence Judgments." *Journal of General Psychology* 129 (2002): 257-300.

Purpura, James E. *Assessing Grammar*. Cambridge: Cambridge UP, 2004.

Rodriguez, Michael C. "Construct Equivalence of Multiple-choice and Constructed-response Items: A Random Effects Synthesis of Correlations." *Journal of Educational Measurement* 40.2 (2003): 163-84.

Salehi, Mohammad, and S. Bagheri Mohammad. "The Impact of Response Format on Learners' Test Performance of Grammaticality Judgment Tests." *Journal of Basic and Applied Scientific Research* 3.3 (2013): 1335-345.

Schraw, Gregory. "The Effect of Metacognitive Knowledge on Local and Global Monitoring." *Contemporary Educational Psychology* 19.2 (1994): 143-54.

- - -. "The Effect of Generalized Metacognitive Knowledge on Test Performance and Confidence Judgments." *Journal of Experimental Education* 65.2 (1997): 135-46.

Schraw, Gregory, Michael E. Dunkle, Lisa D. Bendixen, and Teresa DeBacker Roedel. "Does a General Monitoring Skill Exists?" *Journal of Educational Psychology* 87 (1995): 433-44.

Shohamy, Elana. "Does the Testing Method Make a Difference? The Case of Reading Comprehension". *Language Testing* 1.2 (1984): 147-70.

Stankov, Lazar. "Complexity, Metacognition, and Fluid Intelligence." *Intelligence* 28 (2000): 121-43.

Stankov, Lazar, and Lee Jihyun. "Confidence and Cognitive Test Performance." *Journal of Educational Psychology* 100.4 (2008): 961-76.

Tsagari, Dina. "Method Effect on Testing Reading Comprehension: How Far Can We Go?" Master Thesis. University of Lancaster, 1984.

Vougiouklis, Thomas. "Enlarged Fundamentally Very Thin Hv-Structures." *Journal of Algebraic Structures and Their Applications* 1.1 (2014): 11-20.

Vougiouklis, Thomas, and Penelope Kambaki-Vougioukli. "On the Use of the Bar". *China-USA Business Review* 10. 6 (2011): 484-89.

Weir, Cyril J., Ivana Vidakovic, and Evelina Galaczi. *Measured Constructs: A History of the Constructs Underlying Cambridge English Language (ESOL) Examinations 1913-2012*. Cambridge: Cambridge UP, 2013.

Wolf, Darlene. "A Comparison of Assessment Tasks Used to Measure FL Reading Comprehension." *Modern Language Journal* 77 (1993): 473-89.

Yule, George, Jerry Yanz, and Atsuko Tsuda. "Investigating Aspects of The Language Learner's Confidence: An Application of the Theory of Signal Detection." *Language Learning* 35. 3 (1985): 473-88.

Zheng, Ying, Liying Cheng, and Don A. Klinger. "Do Test Formats in Reading Comprehension Affect Second-language Students' Test Performance Differently?" *TESL Canada Journal* 25.1 (2007): 65-80.