A Case of the Validity Investigation of Concordance-based Cloze Testing: Construct Relevance Revisited

กรณีศึกษาความตรงของการทดสอบแบบโคลซคำในบริบทคอนคอร์แดนซ์: พิเคราะห์ความเกี่ยวข้องกับความหมายคะแนน

Kunlaphak Kongsuwannakul*
กุลภักดิ์ กองสุวรรณกุล*

สาขาวิชาภาษาต่างประเทศ มหาวิทยาลัยเทคโนโลยีสุรนารี
School of Foreign Languages, Suranaree University of Technology, Thailand

ABSTRACT

Investigating construct validity is a process that is essential especially for devising a new item type. However, the distinction between construct-relevant and construct-irrelevant variances may not always be sufficient. In this article, the construct of the concordance-based cloze item type (henceforth ConCloze) will be defined through two research projects. The first one is the doctoral dissertation (Kongsuwannakul, 2017 Investigating the Construct Validity of a Concordance-based Cloze Test: A Mixed-methods Study), in which the construct validity of the item type is defined with an iterative research design. Item components and a variety of changes to them are used insofar as information about the language processes and domains is obtained out of the data. The other research project is a follow-up study (Kongsuwannakul, 2019 Suranaree University of Technology Students' Language Domains in Engaging with a Concordance-based Cloze Test: A Contrastive Approach), in which a contrastive approach is used for identifying distinguishing language domains in ConCloze. It is found that language domains such as knowledge of lexical semantics are not sufficiently distinguishing the construct of the item type. As for theoretical implications, it will be argued that construct-relevant variance could be divided into construct variance and construct-peripheral variance.

ARTICLE INFO

Article history:
Received 22 March 2020
Received in revised form
22 May 2020
Accepted 25 May 2020
Available online
8 June 2020

Keywords:
Construct relevance (ความเกี่ยวข้องกับความหมายคะแนน), Variance (ความแปรปรวน), Construct definition (การนิยามความหมายคะแนน), Concordance-based cloze (ข้อสอบโคลซคำในบริบทคอนคอร์แดนซ์), Case study (กรณีศึกษา), Revisiting categorization (การพิเคราะห์การแบ่งประเภท)

*ผู้เขียนที่ให้การติดต่อ
E-mail address: kunlaphak@hotmail.com

บทคัดย่อ

การศึกษาความสมเหตุสมผลของความหมายคะแนนเป็นกระบวนการที่จำเป็นอย่างยิ่งในการประดิษฐ์ชนิดข้อสอบชนิดใหม่ขึ้นมา อย่างไรก็ดี ความแตกต่างระหว่างความแปรปรวนที่เกี่ยวข้องกับความหมายคะแนนกับความแปรปรวนที่ไม่เกี่ยวข้องอาจไม่เพียงพอเสมอไป ในบทความนี้ความหมายคะแนนของชนิดข้อสอบแบบโคลซคำในบริบทคอนคอร์แดนซ์ (ConCloze) จะนิยามผ่านสองโครงการวิจัย โครงการแรกเป็นวิทยานิพนธ์ (Kongsuwannakul, 2017 Investigating the Construct Validity of a Concordance-based Cloze Test: A Mixed-methods Study) ซึ่งความตรงความหมายคะแนนได้นิยามผ่านการออกแบบการวิจัยแบบวนซ้ำ องค์ประกอบข้อสอบและการเปลี่ยนแปลงองค์ประกอบใช้เพื่อดึงข้อมูลเกี่ยวกับกระบวนการทางภาษาและโดเมนภาษา อีกโครงการวิจัยเป็นโครงการวิจัยต่อยอด (Kongsuwannakul, 2019 Suranaree University of Technology Students' Language Domains in Engaging with a Concordance-based Cloze Test: A Contrastive Approach) ซึ่งใช้กรอบวิธีวิจัยแบบเปรียบเทียบความต่างเพื่อหาโดเมนภาษาที่มีความจำเพาะกับ ConCloze ผลการศึกษาพบว่า โดเมนภาษาเช่นความรู้อรรถความหมายของคำยังไม่มีความจำเพาะมากพอที่จะเป็นความหมายคะแนนของชนิดข้อสอบ ConCloze ความเกี่ยวพันเชิงทฤษฎีจึงอาจกล่าวได้ว่า ความแปรปรวนที่เกี่ยวข้องกับความหมายคะแนนควรแบ่งออกเป็นความแปรปรวนความหมายคะแนนและความแปรปรวนความหมายคะแนนรอบนอก

Introduction

In educational measurement, an observed score is the sum of a true score and measurement error variance (Feldt & Brennan 1989). Put into practice, the true score often relates to construct-relevant variance, and error variance to construct-irrelevant variance (Messick, 1989). In this article, it will be argued that the distinction between construct-relevant and construct-irrelevant variances may not be sufficient in actual practices of construct validation. For the purpose, construct-relevant variance should be subcategorized into construct variance and construct-peripheral variance, at least as far as language assessment is concerned.

Exploring the distinction concerning construct relevance can be useful for two reasons. The first one is addressing scarcity. The distinction between construct-relevant and construct-irrelevant variances has been around in the literature since at least Cronbach & Meehl's (1955) discussion of variances that are relevant or irrelevant to construct validation. While efforts to deal with the relationship and hierarchy of construct validity are not unheard of (e.g., Embretson, 2007; Lissitz & Samuelsen, 2007; Mislevy, 2007; Moss, 2007), little is investigated about the distinction in the field of language testing. Specifically, the two categories of construct-relevant and construct-irrelevant variances tend to be used as is when it comes to language testing. For example, Reardon, Kalogrides, Fahle, Podolsky, & Zárate (2018), inter alia, investigate the relationship between a test-item format and achievement gaps in English language test performance between males and females. They state explicitly that they are unable to determine if the effect of the format is construct-relevant or construct-irrelevant variance (ibid., p. 285). Accordingly, bringing up the topic in this article may draw attention to the distinction and directly create more conversation about how to best draw the distinction.

The second reason for exploring the classification of construct-relevant variance is that the topic is a foundation for all language-testing practices. According to Messick (1995, p. 743), for instance, the notions of construct-relevant and construct-irrelevant variances are all but essential for any educational measurement. This article aims to propose an alternative way of drawing a distinction in relation to construct-relevant variance. Accordingly, it may be argued that the study efforts can have a far-reaching impact because they are tapping into a foundation area for any language-testing projects.

The structure of this article is as follows. Investigating construct validity in my doctoral dissertation will be first outlined. The influence of the research design will be discussed in the case study, highlighting how the very design of research for construct validation leads to response invalidity and supposedly a remnant of construct irrelevance remaining in the construct definition. Then a follow-up research project will be presented, aiming at dealing with the construct-irrelevant variance. However, in an attempt to do just that, a seemingly perpetual variance will be identified, marking a necessity for another category that should be part of the construct-relevant variance. The article ends with a suggestion for use of two subcategories under construct-relevant variance.

Cutting through iterations of validity investigation

In Kongsuwannakul (2017), the construct validity is investigated for a concordance-based cloze test (henceforth ConCloze). Depicted in Figure 1, the item type has been argued to be an innovative item type because prior to the study, to the best of my knowledge, no one has collected empirical data for validation purposes pertaining to it. The framework for validity investigation is to use different test versions, called ConCloze 1–7, to collect data. In doing so, validity evidence accrues for the validity argument through different test designs (cf. Mislevy, 2007 for the concept).

1 by ensuring that they are _____ about ethics issues, have learned
2 and white and more highly _____ and less religiously affiliated than
3 and experience to make a[n] _____ decision about what they need
4 contributing to negative stereotypes of _____ girls. In this section, symbolic
5 a city (such as a[n] _____ labor force, low taxes, and
6 and the whiff of gunpowder. _____ society picks over the descriptions
7 tended to be Caucasian, highly _____, upper income status, and from

All the lines above miss the same word. Which of the following should be that word?

A articulate **B** cultured **C** educated **D** scientific

Figure 1 A ConCloze item (Kongsuwannakul, 2017, p. 341)

In Kongsuwannakul's (2017) ConCloze 1, a prototyping version is created based on the test specification of the study, and quantitative data is collected from a small sample of test takers, 38 respondents (for more information regarding test writing, cf. Kongsuwannakul, 2017, pp. 69–85). Systematic patterns of item difficulty and discrimination, and of test

reliability (Cronbach's alpha coefficient of 0.84) are investigated, allowing an inference that the item type is likely to tap into an underlying domain of competence. In ConCloze 2–4, another prototyping version is created, with modifications made to a few items taken from ConCloze 1. They are called item variants (IVs). Verbal reports are collected from a small sample of respondents, 12 in total. Patterns of substantive processing during task engagement are found in analyzing verbal reports of one item selected purposefully. This allows inferences: testing compatibility of a given word in context is a key language process tapped into when dealing with the item type, and the words in the concordance lines and the options are likely processed for task completion. Regarding test usability, observations during the respondents' engagements with the test tasks, five each, and analyzing the verbal reports allow an interpretation: the IVs and hence the item type itself are usable.

In ConCloze 5, a field-test version is created out of ConCloze 1 items. The item components are modified, and quantitative data is collected from a larger sample of test takers, 285 respondents. In analyzing different IVs, patterns are found, allowing an inference that semantic distance among the item options is a key difficulty driver. In ConCloze 6, a version for operational use is created on the basis of word frequency level for the options. A criterion test, Read's (1998, cited in Cobb, ca. 2011) word association test (WAT), is also administered in this version. Item responses are collected from a total of 247 test takers. Key findings are that word frequency of the options does have an impact on item difficulty, and knowledge of lexical semantics and knowledge of word association are tapped into by ConCloze.

In the final ConCloze version, ConCloze 7, all verbal reports from ConCloze 2–4 are analyzed as part of an effort to fine-tune the research findings derived from ConCloze 1–6. It is found that out of 60 verbal reports, testing a meaningful compatibility of a word in context is the prime process during ConCloze engagement. Lexical semantic knowledge and knowledge of word association are also supported in this version to be primary language domains assessed by the item type.

The test versions outlined above are designed so that modifications either to test-task content or to item formats lead to observable changes in the patterns of item responses and verbal reports. For example, all items in ConCloze 1 have seven concordance lines in each test task whereas some items in ConCloze 5 have ten concordance lines. The item responses of this IV in ConCloze 5 are used for investigating the effects of more concordance lines on item difficulty and thus substantive processing in the item type. Figure 2 shows the conceptual framework of validity investigation through this iterative design, in which construct-related evidence is collected, and generalizability to the universe of admissible observations becomes greater through increasing facets of observations. Therefore, out of the test versions, construct-relevant processes and domains are obtained and refined through iterations.

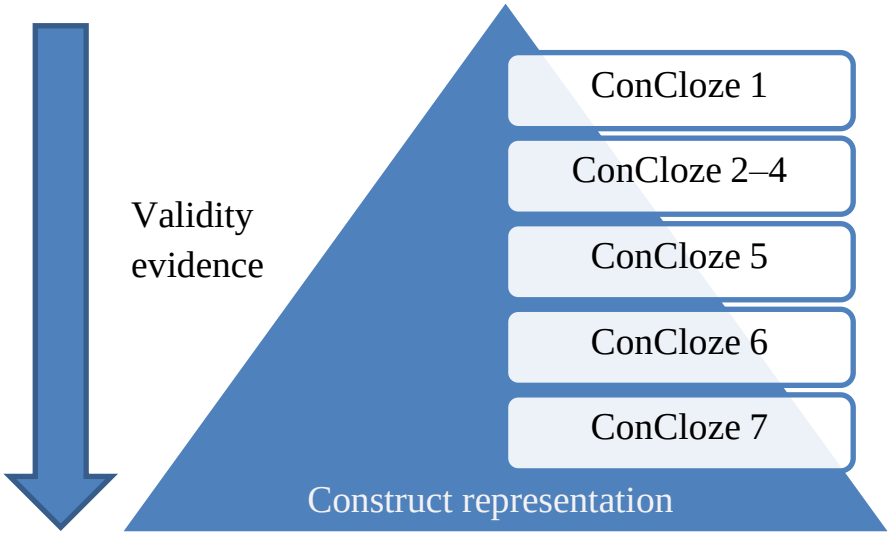

Figure 2 Collecting validity evidence through ConCloze test versions

Despite accruing facets of evidence for generalizability validity, response invalidity persists to some extent in the validation process. A weakness of the dissertation is the voluntary nature of all responses and task verbalizations. On the one hand, gathering test responses from volunteers complies with general research ethics. The ethics require the research data to be obtained on a voluntary basis. Respondents must by no means be coerced to give responses. Yet, on the other hand, doing so for research data could mean that most respondents must have some motivation in completing the whole test tasks. For example, in ConCloze 6, where the WAT is also administered alongside ConCloze, there are altogether 54 test items (30 WAT items + 24 ConCloze items) that the test takers have to complete. Given such a long test battery, participating in research on a voluntary basis may generally indicate that they are likely to be highly motivated. Then, highly motivated students tend to have a high proficiency too (Schmitt, Dörnyei, Adolphs, & Durow, 2004). However, in proportion, there should be more test takers with moderate proficiency than those with a high proficiency. Accordingly, my doctoral dissertation potentially has an epistemological flaw in that the responses may not reflect the ESL/EFL population proportionately. This is thus an inherent response invalidity and presumably construct-irrelevant variance remaining in ConCloze construct definition.

Sieving construct domains out of contrastive pairs

In Kongsuwannakul (2017), convenience and snowball samplings are used for collecting qualitative data, especially verbal reports. These sampling methods have a drawback in that only those who can verbalize their thoughts in English choose to participate in the study. An inference could thus be that the qualitative findings of the dissertation may not fully represent both highly-proficient and weak test takers in the population of academic English non-native users. This is so because those prospective test takers who have very low English proficiency would not choose to participate in the first place and so would be effectively barred from the study given that they could not produce intelligible verbal reports in English.

Given the flaw in the first empirical ConCloze study (Kongsuwannakul, 2017), a follow-up study is initiated, namely Kongsuwannakul (2019). It is designed so that prospective verbalizers with very low proficiency can think aloud their task engagement in their mother tongue, Thai. These verbalizations are then contrasted with the verbal reports of proficient test takers. The aim is to see what language domains those highly proficient respondents use specifically in dealing with ConCloze test tasks. This is represented by area no. 1 in Figure 3.

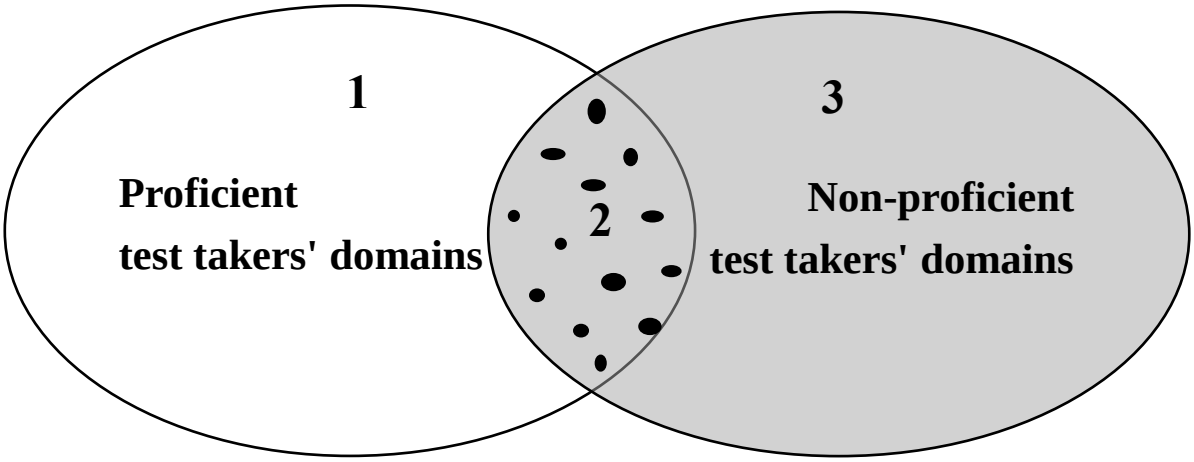

Figure 3 Investigation concept for proficient and weak test takers
(Kongsuwannakul, 2019, p. 4)

In Kongsuwannakul (2019), four proficient respondents and four weak respondents are the sample of the study. They are selected based on their grade results of foundation English courses at the university. These grade results function as surrogates for their proficiency in English. Their verbal reports are transcribed and annotated for pattern investigation of underlying processes and language domains that the proficient groups and non-proficient groups use for task completion. Figure 4 presents a snapshot when a respondent is engaging in one of the test tasks, and Figure 5 the results of the verbalization analysis.

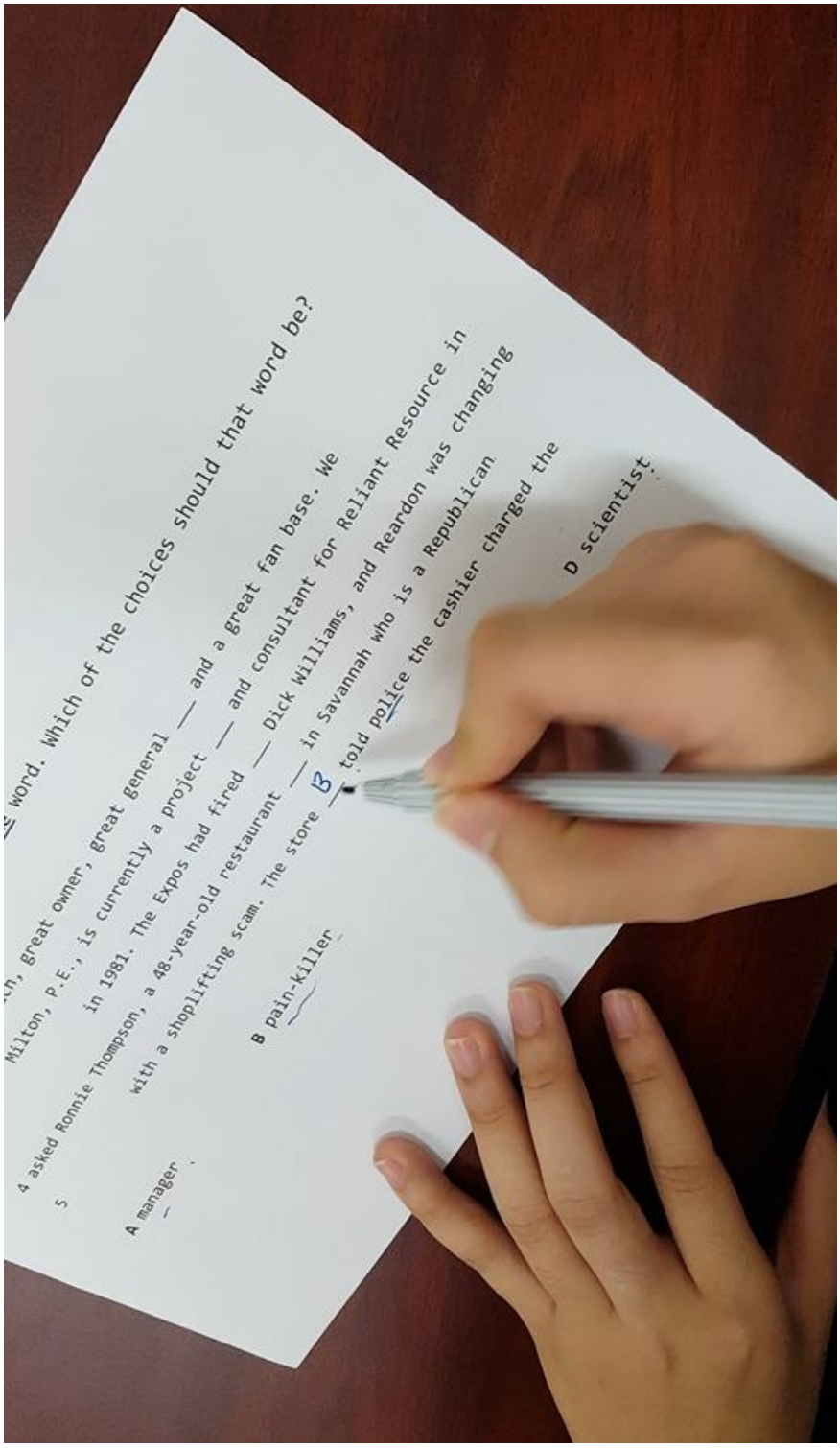

Figure 4 ConCloze test task engagement (Kongsuwannakul, 2019, p. 47)

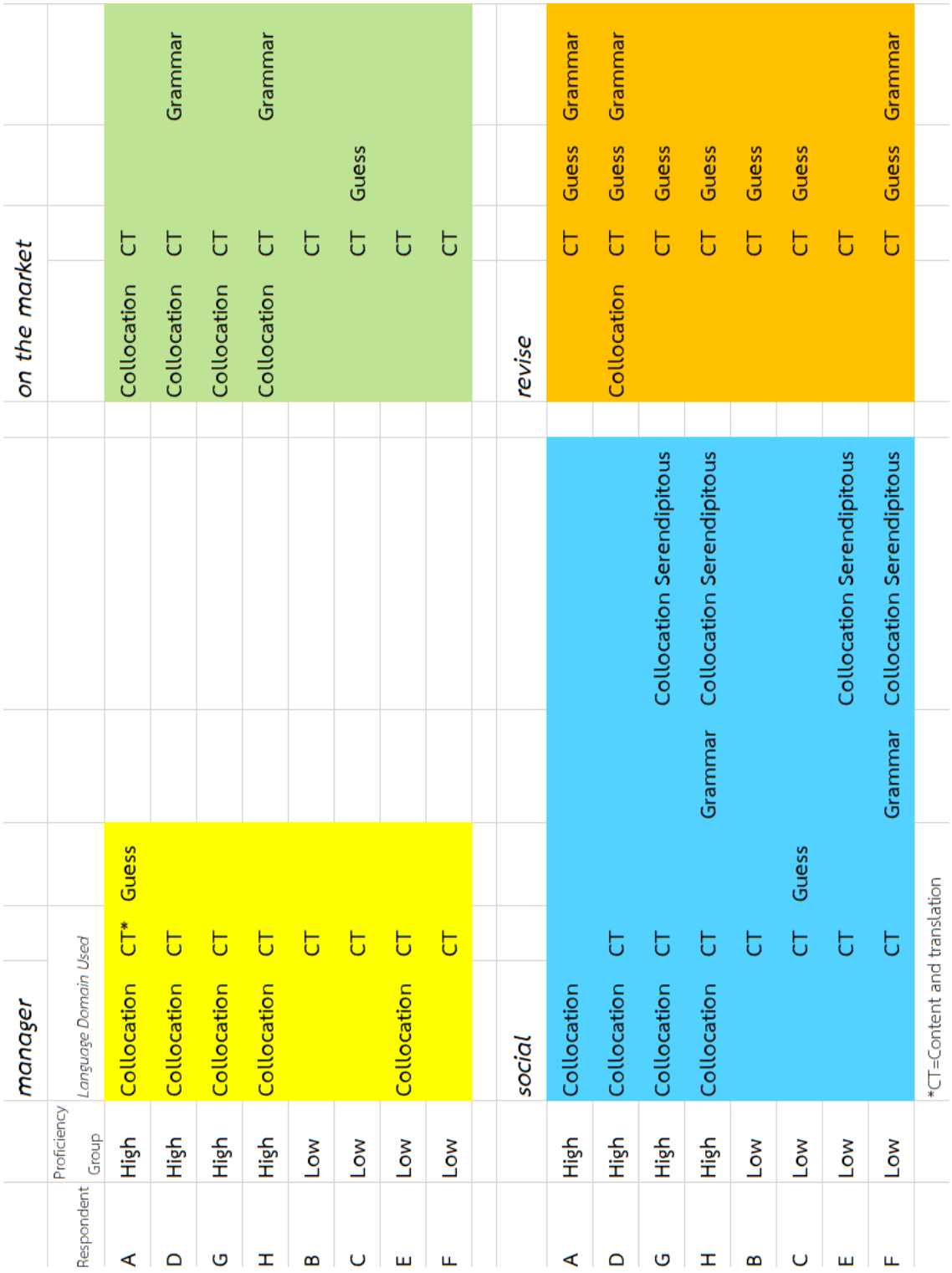

Figure 5 Language domains in ConCloze engagement

In Figure 5, the domain that all proficient respondents (respondents A, D, G, and H) tend to use when dealing with ConCloze tasks is collocation. The knowledge of collocation can thus be said to be a distinguishing language domain that differentiate proficient test takers of ConCloze from those that are weak. Also in Figure 5, the domain of content and translation is found to be the domain that both proficient and weak respondents use almost unanimously during ConCloze engagement. Accordingly, the domain cannot be categorized as a distinguishing domain that separates proficient and weak respondents. However, given that the domain is used for task engagement, it has to be considered construct-relevant to the ConCloze construct. This is equal to area no. 2 in Figure 3, which represents the domains that both of the proficiency groups use for task engagement. In other words, the domain of content and translation cannot be assigned downright to construct-irrelevant variance, because all the test takers still use the domain for completing the test tasks.

In passing, one distinct domain that appears in one item only of Figure 5 should be mentioned for clarification. It is collocation serendipitous that is created in a Grounded Theory-oriented way specifically for the phenomenon observed in the item *social*. Figure 6 depicts a respondent engaging in the test task, with the phrase 'a sense of' in concordance line 2 underlined for being a clue in the concordance-based context. From an immediate retrospective interview, the respondent (respondent F) seeks to connect the phrase with option A 'common' because the fixed phrase 'common sense' is known to the respondent. On the one hand, the occurrence of the domain *collocation serendipitous* is not consistent across items or proficiency groups, as displayed in Figure 5, and so the domain would have to be classified as construct-irrelevant variance. On the other hand, however, observing this phenomenon and annotating the verbal reports with the domain collocation serendipitous help to highlight the validity of the domain of collocation. The reason for this is that fixed phrases such as 'common sense' are part of collocation (cf., e.g., Sinclair, 2004). In other words, the very finding that the respondents seek to meaningfully connect the clues in the concordance lines with the options is a piece of substantive validity evidence for the ConCloze construct representation.

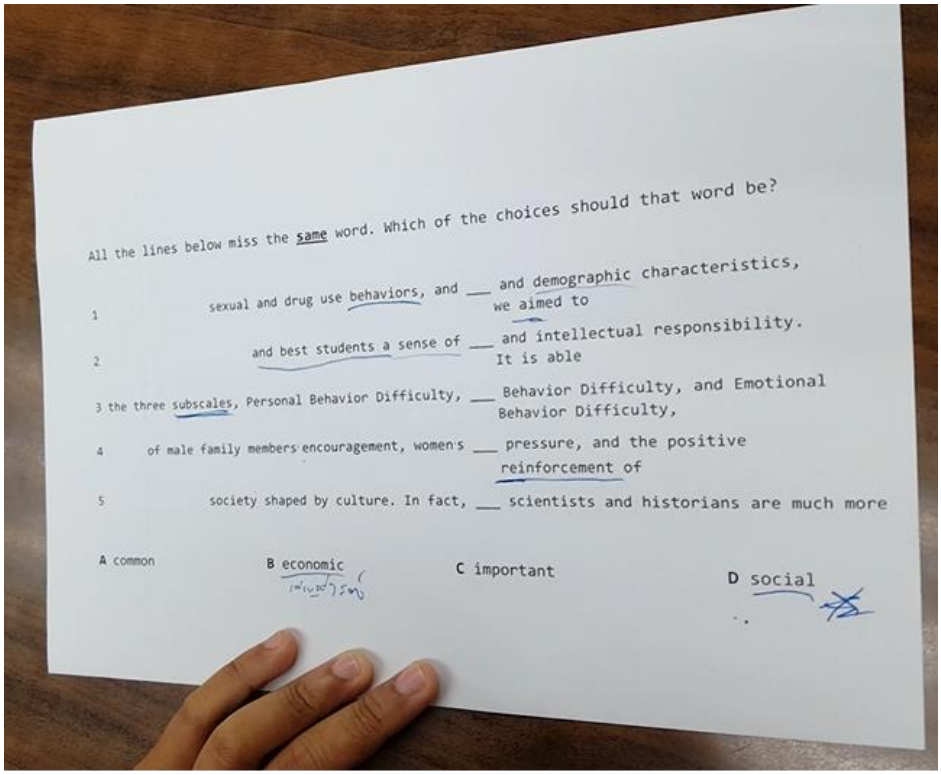

Figure 6 Marking clues during ConCloze engagement (Kongsuwannakul, 2019)

Discussion and implication

Construct-relevant variance is traditionally ascribed to the domains that are used for task engagement (Messick, 1989). In light of the results presented in the previous section, it seems that the notion of construct-relevant variance may be insufficient for differentiating the domain of collocation from the domain of content and translation. On the one hand, both of the domains are found mobilized in the ConCloze verbalizations and so it would be straightforward to simply label them with the category of construct-relevant variance. Yet, on the other hand, when the construct definition is too broad, it means that construct-irrelevant variance is included (Messick, 1995). Including the domain of content and translation straight away into construct-relevant variance would thus make the construct definition of the ConCloze item type too broad and indiscriminatory for different proficiency groups. Given this, it may be argued that the notion of construct-relevant variance should be subcategorized into construct variance and construct-peripheral variance. This then would enable the domain of content and translation to be involved with construct-peripheral variance instead.

Based on the Firthian school, knowing no grammar would allow language learners to say something, but knowing no vocabulary would allow them to say nothing (e.g., Lewis, 2000). The findings in the follow-up study, therefore, reflect this. Words and their meanings in the concordance lines have to be processed in ConCloze and similarly in all types of language test items. Accordingly, the domain of content and translation might be said to be represented by default. For example, Javidanmehr and Sarab (2019) explore the cognitive diagnostic model for a university entrance examination. They discover that in the reading part, vocabulary knowledge and making inferences are among the construct-related areas of competence tested. In light of the subcategorization of construct-relevant variance presented here, construct-peripheral variance could instead be assigned to the domain of vocabulary knowledge in their model and construct variance to the process of making inferences. This is so because the knowledge of vocabulary is anyway invoked during any reading comprehension task.

Stated in the introduction, the notion of construct-relevant variance is related to all language-testing projects. Accordingly, the implication of subcategorizing the variance as proposed in this article can be huge. For example, research on construct definition and construct-validity investigations may need to revisit the categorization of construct-related domains and processes. As far as vocabulary tests are not concerned, chances are the domain of vocabulary and vocabulary knowledge may become demoted from construct-relevant variance to merely construct-peripheral variance. In doing so, construct representation would become sharper as the cognitive domain such as vocabulary knowledge that may not be central or specific to the construct might be trimmed off the construct variance.

Normally, construct definition involves a variety of ways of collecting data and making inferences out of data analyses. On a broader scale, subcategorizing construct-relevant variance may help to define a test construct more meaningfully. Rather than grouping and separating the domains to merely either construct-relevant or construct-irrelevant variance, an investigation can be more focused and selective. The substantive processes and content domains that are used for task completion may be construct-relevant variance, but those that are used specifically by proficient test takers deserve a central role in the construct being defined, and thus construct variance. For example, a cloze test is usually associated with expectancy grammar, in which comprehension of local clues and understanding meaning of words are construct-relevant (Oller 1979). Based on the implication in this article, understanding meaning of words can thus be demoted in such validity investigations.

Acknowledgements

I would like to thank my fund provider, Suranaree University of Technology, for the grant number SUT2-203-61-12-06.

References

Cobb, T. (ca. 2011). **Frequency Based Vocabulary Tests**. Montreal: Tom Cobb; Université du Québec à Montréal. Available: http://www.lextutor.ca/tests/.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. **Psychological Bulletin**. 52(4): 281–302.

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? **Educational Researcher**. 36(8): 449–455.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), The American Council on Education/Macmillan Series on Higher Education. **Educational measurement** (pp. 105–146). Macmillan Publishing Co, Inc; American Council on Education.

Javidanmehr, Z., & Sarab, M. R. A. (2019). Retrofitting Non-Diagnostic Reading Comprehension Assessment: Application of the G-DINA Model to A High Stakes Reading Comprehension Test. **Language Assessment Quarterly**. 16(3): 294–311. doi:10.1080/15434303.2019.1654479

Kongsuwannakul, K. (2019). **Suranaree University of Technology Students' Language Domains in Engaging with a Concordance-based Cloze Test (ConCloze): A Contrastive Approach**. (In Thai). Research report no. X2-203-61-12-06. X Province: X University.

Kongsuwannakul, K. (2017). **Investigating the Construct Validity of a Concordance-based Cloze Test: A Mixed-methods Study**. Ph.D. dissertation, University of Leicester. doi://hdl.handle.net/2381/40396

Lewis, M. (2000). Introduction. In M. Lewis (Ed.), **Teaching Collocation: Further Developments in the Lexical Approach** (pp. 8–9). Massachusetts, USA: Thomson.

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. **Educational Researcher**. 36(8): 437-448. doi:10.3102/0013189X07311286

Messick, S. (1989). Validity. In R. L. Linn (Ed.), **Educational measurement** (3rd ed., pp. 13–103). New York: American Council on Education; Collier Macmillan.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. **American Psychologist**. 50(9): 741–749.

Mislevy, R. J. (2007). Validity by Design. **Educational Researcher**. 36(8): 463–469.

Moss, P. A. (2007). Reconstructing Validity. **Educational Researcher**. 36(8): 470–476.

Oller, J. W., Jr. (1979). **Language Tests at School: A Pragmatic Approach**. London: Longman.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. **Educational Researcher**. 47(5): 284–294. Available: https://doi.org/10.3102/0013189X18762105

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and Acquisition of Formulaic Sequences: A Longitudinal Study. In N. Schmitt (Ed.), **Formulaic Sequences: Acquisition, Processing and Use**. Amsterdam: John Benjamins, 54–86.

Sinclair, J. (2004). **Trust the Text: Language, Corpus and Discourse**. Routledge: London.