Some Observations on the Construct Validity of Psycholinguistic Measures in Foreign Language Learning Research

Mun-Hong Choe[a]
Professor, Chonnam National University, Korea

Abstract

Psycholinguistic measures in the context of foreign language (L2) learning research isolate the phenomenon of interest from irrelevant factors and thereby increase the likelihood of scientific observations. This article is a critical review of the issues concerning the construct validity of some of the traditional measures, focusing specifically on three measurement variables: response time, eye movement, and dysfluency. The advantages and liabilities of several task types are discussed in some detail. Furthermore, two widely recognized problems in assessing and validating constructs, construct misrepresentation and intervention of extraneous variance, are recapitulated. In conclusion, it is suggested that a single measure cannot provide strong evidence for the explanandum in L2 research, and thus the triangulation of methodological tools in both their applications and interpretations should be given priority over efforts to extend their individual validity.

Keywords: Foreign language education, construct validity, psycholinguistic measures, research methods

[a] Corresponding author: Mun-Hong Choe, Professor, Chonnam National University, Dept. of English Education, 77 Yongbong-ro, Buk-gu, Gwangju, 61186, Korea. E-mail: munhong@jnu.ac.kr

Received: 12 Nov 2019, Revised: 19 Nov 2019, Accepted: 27 Nov 2019
http://dx.doi.org/10.34226/gcl.2019.9.6.33

Introduction

The field of foreign language (L2) education aims to describe and explain the processes of language learning under conditions other than those of first language (L1) acquisition. Since many of its theoretical concepts are not directly observable, measurement is used to elicit, observe, and record learners' language-related behaviors. Thus, one of the most contentious issues in the field has been assessment measures and their construct validity. Constructs are psychological entities in theory inferred from the interactions of measurable variables, and the construct validity of a task refers to the extent to which one can make inferences about certain hypothesized constructs measured by the task.

Theoretical accounts differ according to the ways in which L2 learning is defined and the types of evidence that can be brought to support them. There are in turn different measurement tasks vis-á-vis different theoretical standpoints. More specifically, they vary in terms of the extent to which (1) constructs are defined in measurable ways, (2) measurement instruments and procedures are systematically developed and implemented, (3) measurement practices are subject to validity evaluation, and (4) reporting of research is adequate for replication and knowledge accumulation (Bachman & Palmer, 2010). Measurement is a data- and theory-driven undertaking (Messick, 1996). That is, interpretations to be made are scientifically defined and the kinds of data to be accepted as evidence for such interpretations are specified. The first is treated under the notion of construct definition, and the second concerns the nature of measurement data. Construct definition explicitly provides intended interpretations to be made on the basis of a measure. Without construct definition, data will be meaningless.

Measurement data are composed of repeated observations of particular behavioral patterns, and these observations are condensed into scores of some kind, which can be conceived as a summary of observed consistencies on assessment devices (Chapelle, 2012). Psycholinguistic measures in L2 research attempt to isolate the phenomena of interest from irrelevant contexts, and hence increase the opportunities for systematic observations. The goal of this working paper is to explicate some of the issues concerning the construct validity of psycholinguistic measures in L2 research.

The validity of psycholinguistic approaches in foreign language education

The most widely cited argumentation against psycholinguistic approaches was sparked by Firth and Wagner (1997), who called for a reconceptualization of L2 research and practice by enlarging its boundary (Lafford, 2007). According to them, L2 education is concerned with the nature of additional language learning and the development of linguistic knowledge and skills, along with such fundamental concepts as multilinguality, language socialization, variability, foreignness and nativeness. As such, it is part of the nexus of approaches to the

wider, interdisciplinary study of language, discourse, and social interaction. However, theories, methodologies, and foci within the field reflect an imbalance between cognitive-mentalistic orientations and social-contextual orientations to language, with the former orientation having been in the ascendancy.

They argue that this has resulted in a skewed perspective on discourse and communication, which considers L2 speakers as defective communicators struggling to overcome an underdeveloped L2 competence, striving to reach the competence of an idealized native speaker. Thus, L2 education should adopt a holistic approach to and outlook on language learning and teaching. Such an approach problematizes and explores the conventional dichotomy between social and psychological approaches to language use and learning. It attends to the dynamics as well as the outcome of language learning and is more emically and interactionally attuned. With such changes in place, the field of L2 education has the capacity to become a theoretically and methodologically richer enterprise.

The majority of researchers and practitioners assume that language learning is essentially psychological, beginning in the mind of an individual. In the process of learning, learners build up their skills piece by piece, moving from the smaller, simpler components to the higher and more complex ones. The more components the learner has, the better the learners are able to attain communicative skills. However, from a sociocultural perspective, the process of acquisition originates in socially constituted communicative practices. That is, even novice learners can become more competent and creative speakers through scaffolded language experiences. Learning is determined therefore by the quality and quantity of opportunities available to the learner. In this respect, any understanding of the development of individual learners and of the psychological mechanisms through which they comprehend and produce the target language should be assessed within their actual learning contexts.

Nevertheless, L2 learning is essentially based on internal processes such as perceiving, remembering, and thinking (Doughty & Long, 2003). The ultimate goal of L2 learning research is to understand how changes in the learners' internal representation of knowledge is achieved and why the changes are sometimes facilitated by environmental and pedagogical support or appear to cease. Given that L2 researchers and practitioners strive to understand mental processes and changing representations of learners' interlanguage, psychological variables are inevitably and justifiably a central focus. Social factors are important, but less controllable in both naturalistic and classroom settings. Even though both psycholinguistic and sociolinguistic approaches are important, the former should be primary since the basic processes of learning need to be described before or alongside the contextual factors that may influence the processes.

With regards to the negative view of theoretical predilections or methodological practices of psycholinguistic measures, research inevitably requires a certain theoretical perspective on the data to describe and explain it. It forces researchers to be explicit about what they consider to be the most relevant features of the data (Felser, 2005; Poulisse, 1997).

Quantifying also helps to give an objective picture of a particular phenomenon, via which researchers can draw a concrete conclusion on what needs to be explained. Experimental measurement needs to be complemented by naturalistic research, but once a theory that yields precise hypotheses has been developed, it is a very efficient way to test them, because they allow researchers to control contextual and situational dimensions that so often blur the results of naturalistic research.

Moreover, it is difficult to conduct empirical research in the environments where foreign languages are learned (e.g., at home, at work, or in the classroom), because of the great number of potentially interfering variables in natural environments. Since it is almost impossible to keep all these variables constant, the results of studies conducted in natural settings often face much disagreement. It is thus reasonable to abstract away from real classroom situations and to conduct empirical research in a controlled setting where intervening variables can be more precisely manipulated. However, it is worth noting that if the desired response is stipulated to subjects in advance, the measurement risks distorting or circumventing the underlying processes. Conversely, if the response is not stipulated, something other than the targeted response may be produced, with the consequence that the data have uncertain bearing on the constructs of interest.

The validity of psycholinguistic measures: A critical review

In this section, three major paradigms of measurement in L2 psycholinguistic research are discussed with a critical view on their construct validity.

Measuring response time

Self-paced reading tasks

Self-paced reading tasks are based on the assumption that the reader reads a passage at a pace that matches the internal comprehension process and therefore an analysis of the reading rate will unearth the comprehension process itself (Tremblay et al., 2011). The interpretation of reading time is actually based on two additional hypotheses: the immediacy hypothesis and the eye-mind hypothesis. The former assumes that the reader tries to comprehend a unit of text (most often, words) as soon as possible rather than waiting until the end of the sentence of which it is part, while the latter assumes that the mind processes the word currently fixated by the eye. In other words, there is no delay between the word being fixated on and the mental processes assigned to that word.

Critics of this method concede that reading time reflects changes in the processing load. Importantly, though, it does not reveal the source of the changes in processing (Ferreira & Yang, 2019; VanPatten, Keating, & Leeser, 2012). Rather, the reading time for a word or a

sentence is the result of several components, and it is unclear how to distribute the aggregate time to these sub-processes. Even if the processes were known, there would be a range of possible theories about the mechanisms in which they function in concert with time. There also exist other interpretational problems in reading-time data; one problem is that the immediacy assumption may not always hold. Readers may continue to process a previous unit of text while inspecting a new one. They may also preview a text segment and begin processing it prior to fully fixating on it. One type of validation is found in the studies examining the correlation between reading time and other measures of comprehension difficulty (Su & Davison, 2019). For example, reading passages are rated for their familiarity, readability, and narrativity, and these ratings are used as predictors of reading time.

Key-press tasks

In these tasks, a text appears on a computer screen and the reader is exposed to successive segments of the text (windows) by pressing a key. The intervals between presses are defined as the reading time for the window. The window may display the entire text or segmented words, phrases, or sentences. For example, in a moving window task, the subject views a computer screen filled with patterns of dashes and spaces in place of text; dashes correspond to letters, and spaces correspond to those between words. Successive words are revealed by pressing a key. With each subsequent key press, the previous word is masked and a new word appears.

The window method is comparable to the eye-tracking method except for regressive eye movements. At the word level, it is sensitive enough to detect frequency effects and at the sentence level, clausal and structural effects have been uncovered. However, critics point out the quantitative and qualitative differences between eye fixations and reading times. According to Clifton et al. (2007), the correlation coefficient between reading times and gaze durations is not particularly high ($r<.60$), and reading times are about 80% longer in moving window tasks than in eye-tracking situations, giving rise to the possibility that this task involves additional processes.

Readers cannot press the response key as fast as their eyes move. Thus, they tend to press the key before they fully process the given input, retaining it in short-term memory, until they find an opportunity to pause and comprehend, usually at the end of a sentence. In addition, the moving window method does not permit regressive movements. So the assumption that it is capable of measuring the changing mental load during reading has been called into question. More recently, researchers use additional techniques to observe discrete testing episodes. The aim is to assess the transient activation or the workload of the reader. Typically, reading and probing alternate: a subject reads a text, a signal is given to alert them to the test, followed by the test probe, a response is made, feedback on speed and accuracy is given, and then reading resumes.

Decision tasks

Decision tasks call for a speeded decision from the subject in response to a target item; responses include YES/NO, CORRECT/WRONG, NEW/OLD, etc. Lexical decision tasks and item-recognition tasks are the two most widely used ones. Other techniques include same or different judgment, sentence verification, and question-answering tasks. The latency of the target item is compared to that of neutral (non-critical) items. The interpretation of the latency depends on the specific research design. In the simplest case, it is inversely proportional to the activation of the information; the greater the activation, the faster the response. Decision tasks are widely used because they afford researchers' control in the choice, placement, and timing of probes, yielding a harvest of findings ranging from lexical access to text-level inference processes.

In fact, decision tasks have a number of advantages over reaction time methods in detecting activation and tracking its time course. Since decision methods permit greater control over the testing situation, they provide an opportunity to monitor the dynamic changes in activation. This can be done by placement of probes in different locations in the text as well as by presenting successive probes in the priming paradigm (Trofimovich & McDonough, 2011). In the priming paradigm, the subject reads sentences and then sees consecutive pairs of test items. The first member of the pair is referred to as the prime and the second as the target. The prime is assumed to make contact with its representation in memory, activate it, and spread activation to other associate concepts in the representation. Facilitation is reflected by a shorter latency for critical items as compared to neutral items, whereas inhibition is observed when context and target stimuli are based on different representations.

However, decision tasks are not without problems. One problem relates to the notion of activation and another to the strategies that participants are likely to use in decision making. Critics argue that it is difficult to distinguish between activation and memory strength (Mueller, Dunlosky, & Tauber, 2016). Memory strength refers to the relative permanence of information over time. A representation like one's name or phone number has been repeatedly strengthened through multiple encodings and thus has a strong trace. Representations are weak if they have not been recently rehearsed. Activation strength refers to the transitory retrieval state of the target; any information can be temporarily active irrespective of its memory strength.

In addition, the idea that spreading activation can be measured has been questioned on the grounds that activation may dissipate so quickly that conventional online methods cannot detect it properly. Ideally, a decision task should be non-intrusive and assess reading processes in their normal state. The task may introduce extraneous mental operations in its implementation. For example, participants' strategies to enhance response accuracy can be problematic. They have to shift repeatedly from comprehending the given text to responding to a target. Each of these tasks is resource demanding and may therefore interfere with one another (Garrod & Pickering, 2016). In short, decision tasks may engender possible

confounds between reading processes that produce activation and the testing context. The typical way of dealing with such confounds is to use controls designed to eliminate unwanted effects.

Naming tasks

In naming tasks, participants are presented with a text followed by the visual appearance of a target, to which they make an oral response such as naming the target item or giving a one-word answer. Naming tasks are based on activation theory, assuming that highly active concepts are more available for utterance and thus positive targets are named more quickly. Because naming does not involve decisions, the criticisms raised about decision tasks can be avoided. They also have advantages regarding the criterion of naturalness; pronouncing a word is more natural to L2 speakers than deciding, for example, whether a target is an extant word or not. The frequency of errors is relatively small and naming assesses availability in working memory as opposed to strength in long-term memory.

Naming tasks are often used in the study of word, sentence, and text processing. Research results may not converge since researchers use different experimental designs and materials. At the word level, the effect of word frequency has been observed. Naming is facilitated with an associative priming. At the sentence level, response latency is a function of the structural complexity of the given sentence; the more complex the underlying structure of the sentence is, the longer it takes speakers to initiate utterance (Hoedemaker & Meyer, 2019; MacDonald, Montag, & Gennari, 2016). At the text level, naming tasks have been used quite successfully to investigate a variety of discoursal processes such as anaphoric reference, conceptual centrality, and elaborative inference. Access to an anaphoric reference depends on its recency and membership in the same category of referents (Koornneef & Reuland, 2016; Smith & Federmeier, 2019). Central concepts in a text are more accessible than peripheral ones, presumably because they are more interconnected with the remainder of the text. The pattern of naming latency also supports the hypothesis of Mckoon and Ratcliff (2018) that readers tend to make inferences to a minimum extent.

However, some critics argue that naming does not imply lexical access (Hino & Lupker, 1996; Morsella & Miozzo, 2002). Rather, both lexical and non-lexical sources of information race to provide a response. When stimulus items are morphologically regular, participants may utilize a non-lexical process of grapheme-to-phoneme translation rules to complete the task. In this condition, naming latency does not reflect access to the lexicon at all. They rather indicate a direct mapping of graphic input into phonetic output. A similar argument is found in the claim that naming latency partly reflects the translation of a phonological input code to an articulatory output code without intermediate lexical processing. In defense of naming tasks, Monsell (2003) notes that although lexical effects may be masked by the output processes that are carried out in parallel with lexical access, the magnitude of lexical effects in naming are in fact comparable to those observed in other

recognition tasks.

Tracking eye movement

Another popular methodological paradigm is eye-tracking. Assuming that eye movements reflect the person's cognitive processing, researchers record language speakers' eye movements and fixations in an effort to make inferences about language processing. Eyes do not move smoothly and continuously. There are ballistic movements called saccades, usually from left to right, fixations and regressions to previous locations in the text. Several measures of eye fixations are available for each region on the screen, including the duration of the first fixation, the sum of the total fixations excluding regressions, and the aggregate duration of all fixations and regressions. The latter is referred to as gaze duration. For example, Hyona, Yan, and Vainio (2018) investigated the effect of word structure on eye-movement patterns, reporting that reading was facilitated when readers were able to preview the initial letters of the word to the immediate right of their current fixation. Using gaze durations, Crossley et al. (2017) investigated the effects of word length and familiarity on processing. Patterns of eye fixations and movements reveal syntactic processing at clause boundaries and in structurally ambiguous regions. Longer gaze durations were reported at clause-final words, especially at sentence boundaries. The point of contention has been to what extent this is true for different types of processing and which aspects of eye movement are informative or necessary for successful processing.

According to critics, eye-tracking is problematic because there are different fixation measures. For example, it is unclear whether researchers sum up fixations for a given word or whether they include return gazes in an aggregate measure. If fixations are not summed up for a word, it is not clear which fixation to use and what to do if each produces a different interpretation (Clifton et al., 2016). In addition, researchers using eye-tracking methods do not analyze processing during saccades that contribute up to 15% of movement (Kieras & Just, 2018). Eye-tracking researchers believe that processing a previous word or sentence while inspecting or previewing a text prior to fully fixating on it is a naturally occurring process and should therefore be given an account. Thus, the most appropriate measure is probably using several measures together.

Recording dysfluency

Dysfluency such as hesitations, false starts, filled pauses, and repetitions, refers to interruptions to a natural flow of speech that are distinguished from speech errors. The study of dysfluency deals not only with the features of dysfluency but with the durations of various grammatical junctures that naturally punctuate speech. Dysfluency has various causes. It may reflect transient increases in processing load, advance planning, and retrieval for an upcoming structural unit, or delay created by the momentary inaccessibility of a needed piece of information. That speakers often pause before and while speech can be explained

in at least three different ways: (1) a tendency to reconsider what to say before actually saying it, (2) time used to plan the details of the next utterance, or (3) a delay in the retrieval of words or other units of utterance (Carroll, 2008). Of these possibilities, pauses for planning the next utterance has received the most concerted attention with the aim of identifying the locus and scope of forward preparation in speech by virtue of examining how far ahead speakers plan and where they tend to hesitate while doing the cognitive work needed for an upcoming stretch of speech. Researchers explore the distributions and durations of dysfluency to determine whether single words or larger groups function as minimal planning ranges.

A major validity issue relates to the classification of dysfluency into one or another category. Behind this issue lies the question of whether differences in the surface forms of dysfluency reflect different production problems or different kinds of preparation (conceptual, lexical, or syntactic). There are reasons to doubt the assumption that dysfluency occurs from a set of identifiable sources. Pauses are more likely to occur in planned than in unplanned speech, whereas false starts, corrections, and repetitions are more likely to occur in unplanned speech. Silent pauses fill a substantial portion of spontaneous speech, between 40% and 50% of speaking time on average, but its occurrence does not necessarily indicate uncertainty or mental difficulty on the part of the speaker (Ellis & Beattie, 2017). Some pauses mark linguistic boundaries; other pauses may be used for purely stylistic effects or aids to the hearer's understanding. Carroll (2008) also pointed out that the causes of any particular dysfluency are by no means obvious as it involves a multitude of factors in combination such as breathing needs, pragmatic intentions, grammatical junctures, or planning problems.

Validating a measure: General discussion

It has been acknowledged that the major threat to the construct validity of psycholinguistic measures is of two types: construct misrepresentation and intervention of extraneous variance. The former indicates the degree to which a measure fails to capture important aspects of the targeted construct, and the latter indicates the degree to which measurement scales are affected by mental and behavioral processes that are extraneous to the targeted construct. The issue of construct misrepresentation usually arises during the conceptualization of a measure, when the link between theoretical interpretations and required evidence is not adequately understood and conveyed into practice. Meanwhile, the issue of extraneous variance is often found during the implementation of a measure, when researchers fail to control or account for the potential influence of the measurement process itself.

Given the extended range of tasks employed by L2 research from discrete-point knowledge items to spontaneous communicative interactions as well as the range of construct interpretations that are based on them, the potential threat to construct validity is

broad and diverse. For example, where interpretations are to be made about the relationship between causal or moderating processes (perception, noticing, cognitive resources of memory and attention, attentional focus, language aptitude, etc.) and L2 outcome, behavioral evidence for such constructs must be specified and associated with measurement tasks. The measures that provide evidence bearing only on the outcome of acquisition (e.g., vocabulary knowledge scales, grammaticality judgment tasks) do little to inform about the variables to which acquisition-related behaviors are ascribed. In other words, interpretations about the relationship between certain causes and outcomes in L2 learning would not be warranted.

For example, Skehan (2018) and Robinson (2005) proposed two contrasting processes that predict similar changes in L2 behavior. Both models predict that the more cognitively complex a task (meaning-oriented communicative activity), the more likely it will yield more complex but less fluent output by learners, under the presumption that task complexity is positively related to L2 learning. However, Robinson argues that the linguistic processing demanded by cognitively more complex tasks entails a mobilization of attention dedicated to language production, and thus pushes the internal system in several ways by fostering linguistic processing that promotes rehearsal in short-term memory and eventual reorganization of form-meaning connections. In contrast, Skehan claims that unmitigated cognitive complexity can have the undesirable effect of overloading the learner's limited attentional resources and hence reinforcing an easy way out through lexical, as opposed to grammatical, processing of L2 input and output. Therefore, during competence-building practice, it is necessary to orchestrate external interventions to make sure that learners consciously attend to the target form and give priority to accuracy. This is an information-processing and skills-acquisition rationale in nature based on the notion of limited cognitive capacity.

The predictions of Robinson will be measurable if one can elicit behaviors that reflect the psychological processes (for example, deeper processing and rehearsal in short-term memory) that reside beyond conscious control. Skehan's theory, on the other hand, requires the measurement of behaviors that reflect meta-cognitive processes (strategic attention and priority decision in this case) should be elicited, which are subject to conscious control. Therefore, each interpretation calls for distinct measures to provide behavioral evidence. Implicit tasks (e.g., decision tasks, naming tasks, eye-tracking methods) will be more appropriate choices for subconscious operations whereas those that reflect language users' moment-by-moment strategic actions (e.g., self-paced reading tasks, window tasks, dysfluency occurrences) will be more relevant to metacognitive operations.

Many L2 studies fail to measure intended constructs because of a mismatch between the assumptions embedded in conventional test development practices and the complex nature of the constructs to be measured. In a review of (quasi-)experimental studies of L2 instruction, Norris and Ortega (2000) compared the observed magnitude judgments (various kinds of grammaticality judgment tasks) and constrained responses (selecting or producing word- or clause-level responses). They argue that constrained response measures reduce

language behavior to single instances of producing a form out of context and thus do not adequately reflect the complexity of interpretations drawn about L2 learning. Given the disjuncture between such isolated language and communicative language use or a learner's underlying representation of L2 knowledge, the link between the measures and the interpretations that can be made about a learner's internal changes is at best tenuous.

Even though behaviors to be elicited in measurement are carefully selected to provide evidence for intended constructs, construct misrepresentation remains a threat unless variable behaviors are specified in analysis. In order to understand what a language task measures, we first have to understand the task-specific processes on which the observed performance depends. Making a valid interpretation on the basis of elicited performance will depend then on understanding to what extent observed behaviors are influenced by the interaction of learner variables with measurement types (Bachman & Cohen, 1998). In L2 research, meaningful measurement scales include amount, duration, frequency, sequences, and comparisons of one sort of behavior with another, and recent empirical approaches to task analysis are helpful in conceptualizing the demands imposed by measurement tasks and the ways in which learners deal with such demands during task performance. In addition, when theoretical conceptions are translated into behavioral measurement, the procedures taken by the researcher may influence results and thereby interpretations. The fundamental question in this implementation stage is to what extent patterns in the behavioral data which are actually elicited, scored, and analyzed can be attributed to the construct, as opposed to irrelevant variance due to measurement error.

The behavior elicitation and observation stage of measurement is particularly susceptible to the occurrence of measurement error. There are a number of factors to be considered for the prevention of procedural inconsistency. Above all, researchers should make sure that all important aspects of the targeted construct are faithfully transformed into measurement tasks and procedures. For example, in the study of developmental sequences in L2 grammatical morphemes, the design of measurement tasks is required to satisfy several requirements to elicit consistent behavioral patterns. Since the initial emergence of a particular morphosyntactic form is implicatively related with the preceding or subsequent emergence of other forms, behavioral data should be gathered across a variety of contexts. Furthermore, given the fact that the initial emergence of a form may occur in different contexts for different individuals, behavioral data should be gathered using a variety of tasks. It is only through the elicitation of extensive amounts and types of L2 behaviors that measurement can show that certain linguistic forms have actually been acquired while other forms have not been acquired yet. If measurement tasks fail to provide the range of contexts necessary for patterns in emergence to be displayed, then interpretations about developmental stages will remain inconclusive (Eskildsen, 2015).

On the other hand, researchers should also be wary of potentially unpredictable sources of measurement error that are associated with directions, environments, individual learners, etc. For example, in a study that aims to make interpretations based on spoken interactions,

interlocutor profiles as well as particular actions undertaken by the interlocutor may influence the learner's performance. Research into oral interview tasks, wherein learners interact with one or more interlocutors, shows that interlocutor variables such as age and gender may influence the amount and quality of language produced by the examinee. In order to reduce the effect of these problems during the elicitation of behaviors, careful pilot-testing and revision of tasks, directions, and administration guidelines are essential.

All in all, it can be said that a single measure cannot provide sufficient evidence for informing the range of interpretations sought in L2 research and that theories which posit certain constructs need to incorporate practical means for observing the full range of the constructs (Rebuschat et al., 2015). Therefore, triangulating the methodological tools in both their applications and interpretations should be given priority over efforts to extend their generalizability. Developing, implementing, and validating a measure cannot be properly accomplished by individual researchers. Only when the recursive, as well as stepwise, phases are tackled cooperatively by the research community will the field be able to make meaningful interpretations about its constructs and truly accumulate knowledge of its own.

References

Bachman, L. F., & Cohen, A. D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research*. Ernst Klett Sprachen.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Carroll, W. D. (2008). *Psychology of language*. Toronto: Thomson Wadsworth.

Chapelle, C. A. (2012). Validity argument for language assessment. *Language Testing*, *29*(1), 19-27.

Clifton Jr, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, *86*, 1-19.

Clifton Jr, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye Movements* (pp. 341-371). NY: Elsevier.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, *54*(5-6), 340-359.

Doughty, C. J., & Long, M. H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology*, *7*(3), 50-80.

Ellis, A. W., & Beattie, G. (2017). *The psychology of language and communication*. NY: Routledge.

Eskildsen, S. W. (2015). What counts as a developmental sequence? Exemplar-based L2 learning of English questions. *Language Learning*, *65*(1), 33-62.

Felser, C. (2005). Experimental psycholinguistic approaches to second language acquisition. *Second Language Research*, *21*(2), 95-97.

Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes*, *56*(7), 485-495.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal*, *81*(3), 285-300.

Garrod, S., & Pickering, M. (2016). *Language processing*. London: Psychology Press.

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1331.

Hoedemaker, R. S., & Meyer, A. S. (2019). Planning and coordination of utterances in a joint naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 732.

Hyönä, J., Yan, M., & Vainio, S. (2018). Morphological structure influences the initial landing position in words during reading Finnish. *Quarterly Journal of Experimental Psychology*, *71*(1), 122-130.

Kieras, D. E., & Just, M. A. (2018). *New methods in reading comprehension research*. NY: Routledge.

Koornneef, A., & Reuland, E. (2016). On the shallow processing (dis-)advantage: Grammar and economy. *Frontiers in Psychology*, *7*, 82.

Lafford, B. A. (2007). Second language acquisition reconceptualized? The impact of Firth and Wagner (1997). *The Modern Language Journal*, *91*, 735-756.

MacDonald, M. C., Montag, J. L., & Gennari, S. P. (2016). Are There Really Syntactic Complexity Effects in Sentence Production? A Reply to Scontras et al. (2015). *Cognitive Science*, *40*(2), 513-518.

McKoon, G., & Ratcliff, R. (2018). Adults with poor reading skills, older adults, and college students: The meanings they understand during reading using a diffusion model analysis. *Journal of Memory and Language*, *102*, 115-129.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241-256.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134-140.

Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 555.

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2016). The effect of identical word pairs on people's metamemory judgments: What are the contributions of processing fluency and beliefs about memory? *The Quarterly Journal of Experimental Psychology*, *69*(4), 781-799.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*(3), 417-528.

Poulisse, N. (1997). Some words in defense of the psycholinguistic approach: A response to Firth and Wagner. *The Modern Language Journal*, *81*(3), 324-328.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*(2), 299-334.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL-International Review of Applied Linguistics in Language Teaching*, *43*(1), 1-32.

Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment*. NY: Routledge.

Smith, C. M., & Federmeier, K. D. (2019). What does "it" mean, anyway? Examining the time course of semantic activation in reference resolution. *Language, Cognition and Neuroscience*, *34*(1), 115-136.

Su, S., & Davison, M. L. (2019). Improving the Predictive Validity of Reading Comprehension Using Response Times of Correct Item Responses. *Applied Measurement in Education*, *32*(2), 166-182.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569-613.

Trofimovich, P., & McDonough, K. (Eds.). (2011). *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics* (Vol. 30). Philadelphia: John Benjamins Publishing.

VanPatten, B., Keating, G. D., & Leeser, M. J. (2012). Missing verbal inflections as a representational problem: Evidence from self-paced reading. *Linguistic Approaches to Bilingualism*, *2*(2), 109-140.

Korean Abstract

외국어 학습 연구에 응용되는 심리언어학적 측정법의 구인타당성에 대한 소고

최문홍 (전남대, 교수)

외국어 학습 연구 분야에서 심리언어학적 측정법은 특정 현상을 무관한 요인들로부터 분리함으로써 과학적 관찰의 가능성을 높인다. 이 연구는 몇몇 전통적으로 흔히 사용되고 있는 방법들의 구성요인 타당성에 관한 주요 사안들을 비판적 관점에서 재고찰한 것으로, 특히 반응 시간, 안구 운동, 비유창성의 세 측정 변인에 초점을 두고 논의하였다. 각 과제 유형이 지닌 장점과 한계점을 상당히 세부적으로 논의하였으며, 목표 구성요인을 평가하고 타당화하는 과정에서 발생할 수 있는 두 가지 널리 알려진 문제인 구성요인 미표상과 외적 분산의 간섭을 적절한 예를 들어 재정리하였다. 결론적으로 어느 한 가지 측정법만으로는 외국어 학습 연구가 설명하고자 하는 바에 대한 충분한 증거를 제공하기 어렵고, 따라서 개개 측정법의 타당성을 확대하려는 노력보다 연구 도구들의 적용과 해석 도출의 과정을 다각화하는 것이 더 중요하다고 제안하였다.

주요어: 외국어 교육, 구성요인 타당성, 심리언어학, 측정, 연구방법