The Inadequate Use of Confirmatory Factor Analysis in Second Language Acquisition Validation Studies

Payman Vafaee[1]
Teachers College, Columbia University
Ilina Kachinske[2]
University American College Skopje

ABSTRACT

The current study aims to demonstrate how the lack of a conceptual framework and the inadequate use of confirmatory factor analysis (CFA) in second language (L2) acquisition validation studies can lead to misconceptions about the nature of data collected via different measurement instruments. To this end, we reanalyzed data from three factor-analytic validation studies on several measures of L2 explicit (EXK) and implicit (IMK) knowledge and demonstrated how an inadequate implementation of CFA in these studies has led to unwarranted validity claims. Following several criteria (e.g., presence or absence of time pressure, or drawing of attention to form or meaning), Ellis and Loewen (2007), Bowles (2011), and Zhang (2015) created test batteries that included different types of tests hypothesized to be *distinct* measures of EXK or IMK. Our re-analysis included the original CFA models retained in these studies, together with new theoretically and empirically plausible alternative models. Results demonstrated that the conclusions reached in the original studies were compromised by the existence of alternative or even equivalent CFA models that fit the date, and the measures included in these batteries were not actually distinct measures of EXK and IMK.

Keywords: *explicit knowledge, confirmatory factor analysis, grammaticality judgment tasks, implicit knowledge, validation and validity*

[1] Payman Vafaee is a Lecturer in Applied Linguistics at Teachers College, Columbia University. He earned his Ph.D. in Second Language Acquisition from the University of Maryland, and his research interests include second language assessment and testing, quantitative research methodology, instructed SLA, and cognitive individual differences in SLA. He has published in several journals such as *Studies in Second Language Acquisition*, *Instructed Second Language Acquisition,* and the *International Journal of Language Testing*. Correspondence should be sent to Payman Vafaee, e-mail: pv2203@tc.columbia.edu

[2] Ilina Kachinske received her Ph.D. in Second Language Acquisition at the University of Maryland. Her research focuses on practical issues concerning the cognitive aspect of second language acquisition such as different implicit and explicit learning mechanisms, knowledge and instruction, cognitive individual differences, their interaction with different types of instructions and impact on L2 attainment, conditions that make explicit instruction beneficial to the acquisition process, as well as the role of declarative and procedural knowledge in the process of automatization. She explores these questions by using both artificial and natural languages. Correspondence should be sent to Ilina Kachinske, e-mail: ilina.stojanovska@gmail.com

INTRODUCTION

Second language acquisition (SLA) researchers need different types of measurement tools to answer research questions or test hypotheses.[3] This has led to an increasing number of studies through which the more commonly used measures are validated. The use of passive voice (i.e., the measures *are* validated) implies that SLA researchers consider validity as an inherent quality of a measure itself, and its definition is whether a measurement tool measures what it is supposed to measure (Chapelle, 2013). However, "[t]o claim that validity refers simply to demonstrating that a 'test measures what it purports to measure' or that it is an inherent property of a test is to ignore at least 70 years of research on validity theory and test validation" (Sireci, 2009, p. 28). More recently, rather than being treated as an inherent quality of tests themselves, validity has been considered as a quality that pertains to test interpretations and uses made based on test scores. In this sense, validation is the process of collecting and evaluating evidence for/against the plausibility and appropriateness of these interpretations and uses (Messick, 1989), and validity is an argument built upon the collected evidence and logical reasoning to justify the interpretations and uses.

To streamline the validation process in this contemporary sense, Kane (2006) proposed an argument-based conceptual framework that guides researchers in deciding what kind of evidence to collect and how to build a validity argument. To enhance the quality of their validation studies, SLA researchers can benefit from this framework because it can help them avoid methodological pitfalls, such as the inadequate use of factor-analytic techniques (examined in the section below) and, consequently, unwarranted validity conclusions. The goal of the current study was to demonstrate how the adoption of Kane's framework, at least in part, could enhance the quality of validation studies in SLA.

To this end, the current study demonstrates how the inadequate use of confirmatory factor analysis (CFA) can lead to unwarranted validity claims about the nature of data collected via commonly used measures in SLA research. For instance, two commonly tested constructs in SLA research are second language (L2) explicit (EXK) and implicit (IMK) knowledge. To be able to measure these two constructs distinctively, several efforts have been made to create and validate their *distinct* measures. Following several criteria (e.g., presence or absence of time pressure, or drawing of attention to form or meaning), Bowles (2011), Ellis and Loewen (2007) and Zhang (2015), among others, created batteries that included tests hypothesized to be distinct measures tapping EXK and IMK. These batteries included metalinguistic knowledge tests (MKTs), timed (TGJTs) and untimed (UGJTs) grammaticality judgment tasks, oral narrative tasks (ONs), and elicited imitation tasks (EIs). The authors claimed that by, for example, applying time pressure on some measures (e.g., TGJT) but not on the others (e.g., UGJT), two distinct types of measures tapping EXK and IMK were created. By testing and retaining two-factor CFA models, they provided support for the claim that the UGJTs and MKTs were measures of EXK and that the TGJTs, ON, and EI were measures of IMK.

By adopting Kane's (2006) validation framework we sought to replicate and extend these studies by including both the original models that the authors retained, as well as theoretically and empirically plausible alternative one-factor CFA models. The support for these alternative models suggests that the validity claim that distinct measures of EXK and IMK were created in the above studies was not empirically warranted. We hope that the current study can help SLA

[3] Throughout the current text, terms such as "measurement tool," "measure," "data collection instrument," and "test" are used interchangeably.

researchers make informed decisions in the process of validation studies, especially in using CFA as a data-analysis method.

BACKGROUND

The Argument-Based Validation Framework and CFA

The 1980s marked the beginning of a series of serious discussions in educational measurement regarding the definition and scope of validity (e.g., Cronbach, 1988). The publication of AERA/APA/NCME standards for educational and psychological testing in 1985 and a seminal paper by Messick entitled “Validity” in 1989 led to the replacement of three different kinds of validity (i.e., content, construct, and criterion-related) with a single unified view of validity—one which considers construct validity as central. In this view, content and correlational analyses were re-conceptualized as methods for probing construct validity (Chapelle, 1999). Additionally, this updated conception of validity was reconsidered as pertaining to interpretations or inferences made based on test scores, rather than the tests themselves.

Building on previous developments, Kane (2006) proposed an argument-based validation framework, which has provided test users with guidance in arguing for or against the validity of interpretations based on test scores. In this validation approach, the aim is to provide an overall evaluation of the intended interpretations derived from test scores, rather than solely examining the qualities of the test itself. This is achieved through a coherent analysis of all of the evidence for and against the proposed interpretations (Cronbach, 1988) by building two types of arguments in two consecutive steps: (1) an *interpretive argument*, which specifies in some detail the proposed interpretations based on test scores, and (2) *a validity argument*, which results from evaluating the overall plausibility of the interpretations as outlined in the interpretive argument (Kane, 2006).

An interpretive argument includes a chain of inferences about the intended interpretations, and for this reason, inference is a key concept in the argument-based validation framework. Kane (2006) defined and used the concept of inference consistent with Toulmin’s (2003) description of informal or practical arguments as used in nonmathematical fields like law.[4]

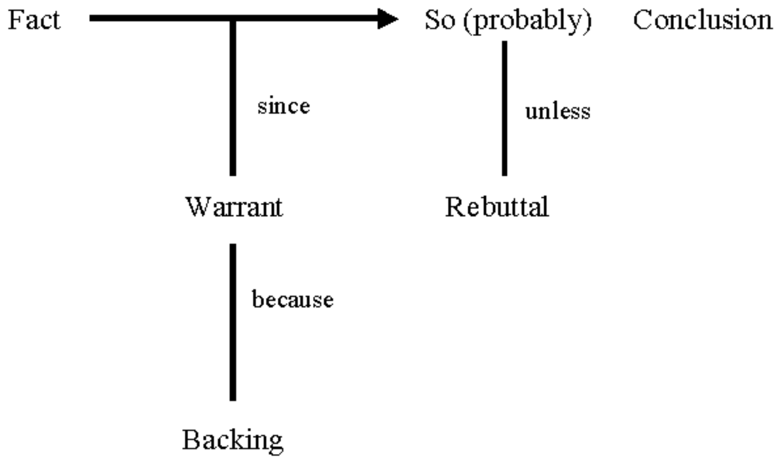
According to Toulmin (2003), an inference is a logical movement from one fact or proposition to another (Chapelle, 2013). Each informal argument has the following structure and components. First, conclusions or interpretations are drawn about, for example, a test’s scores. Such conclusions follow from a chain of reasoning that starts from data or any other empirical observation. In Mislevy, Steinberg, and Almond’s (2003) terms, these conclusions are referred to as *claims*. Claims are made on the basis of data or observations that Toulmin (2003) referred to as *grounds*. The inferential link between a claim and grounds is authorized by a *warrant*, which can be a law, generally held principle, or established procedure. The warrant, in turn, has its own underlying assumptions, and the warrant can license the inferential link only if there is enough support or *backing* for the plausibility of these assumptions, in the form of, for instance, a

[4] The reasoning process in informal arguments is different from that of formal arguments in math, for example, in which premises are taken as given and do not need to be verified (Chapelle, Enright, Jamieson, 2008).

scientific theory and/or well-established empirical evidence. However, even when the plausibility of the underlying assumptions of a warrant is supported with a seemingly indisputable backing, alternative or rival theoretical claims and/or empirical evidence may undermine the plausibility of these assumptions. For this reason, *rebutting* is always required in any validation process.

The rebuttal weakens the inferential link between the claim and grounds by questioning the plausibility of the underlying assumptions of the licensing warrant; and if an inference survives a rebuttal, it creates a strong basis for moving to the next inference in the interpretive argument (Chapelle et al., 2008; Mislevy et al., 2003). Figure 1 depicts the structure of an informal argument as outlined by Toulmin (2003).

FIGURE 1
The structure and components of an informal argument outlined by Toulmin (2003).



As seen in Figure 1, developing a plausible informal argument with warranted conclusions requires the test of potential rebuttals against the supporting evidence for the plausibility of the underlying assumptions of its licensing warrant. Without this piece, the seemingly well-supported conclusions may be invalid because untested alternative or even more plausible explanations may exist. In other words, failing to test the plausible rebuttals may create confirmation bias in the conclusions of an informal argument.

The current study aims to demonstrate how inadequate testing of plausible rebuttals in the process of validation can create confirmation bias in conclusions about the nature of data collected via measurement instruments of second language EXK and IMK. The validation studies on these measures scrutinized in the current study employed CFA to provide backing for the plausibility of the underlying assumptions of one of the warrants that authorized the link between the *explanation* inference and the subsequent inference. The explanation inference, in the chain of other inferences in an interpretive argument, links the expected scores (i.e., reliable, generalizable/dependable scores) to a theoretical construct, thus adding meaning to the test scores. One important warrant for this inference is that the expected scores from a single test are attributable to a single theoretically defined construct. One method of evaluating the plausibility of the underlying assumptions of this warrant entails examining the factor structure of a test or a battery of tests by using factor analytic methods such as CFA.

CFA is a special case of the general family of structural equation modeling, commonly used in many social sciences for investigating theory-derived structural relationships and hypotheses about a set of measured variables (Mueller & Hancock, 2008). Conducting CFA to its full capacity entails several consecutive steps: (1) specifying the hypothesized model, (2) identifying and estimating the model's parameters, (3) assessing the data–model fit, (4) making possible and plausible modifications in the hypothesized model, and (5) identifying and testing alternative models that may rebut or weaken the structural inferences made in the hypothesized model.

Using the informal argument structure, the core component of Kane's (2006) argument-based validation framework, we provided an empirical illustration of the problem of failing to test CFA alternative models (stage 5 above) in validations studies. Failing to test CFA alternative models is equivalent to failing to test rebuttals in an informal argument, without which the argument claims may be unwarranted. In the section below, we give a brief overview of L2 EXK and IMK, the two constructs that were the main topic of research in the studies we reanalyzed.

L2 Explicit and Implicit Knowledge

Second language acquisition involves both explicit and implicit learning mechanisms that lead to both L2 EXK and IMK. We have a conscious awareness of our EXK, but we do not have such an awareness of our IMK (DeKeyser, 2009). One important question regarding these two constructs is whether EXK may become IMK and vice versa. This has led to an ongoing debate in the field of SLA known as the Interface Issue, which entails three alternative views: the non-interface, strong-interface, and weak-interface positions.

The non-interface position claims that L2 explicit and implicit learning are two distinct phenomena resulting in EXK and IMK, respectively, with no interface between these two kinds of L2 knowledge (e.g., Krashen, 1994). On the other hand, the strong-interface position relies on the models of Skill Acquisition Theory, such as ACT-R (Anderson & Lebiere, 1998), and makes a distinction between declarative knowledge (i.e., knowledge that something is) and procedural knowledge (i.e., knowledge how to do something). According to this model, learners first develop declarative knowledge, and as a result of extensive practice, they develop procedural knowledge. Further practice leads to automatized knowledge, which may not require any conscious processing. The strong-interface position implies a causal relationship between declarative knowledge and proceduralized and automatized knowledge (e.g., DeKeyser, 2009).

The weak-interface position claims that EXK only facilities implicit learning and mediates the development of IMK indirectly. For example, EXK can help learners to recognize the gap between the input and the linguistic knowledge they possess (e.g., N. Ellis, 2005).

To test these rival interface positions, SLA researchers need distinct measures of EXK and IMK. To this end, several validation studies have attempted to investigate whether specific data collection instruments (e.g., different kinds of GJTs) can distinctly tap the distinct constructs of L2 EXK and IMK.

An important consideration here is that in any educational and psychological assessment, there are two different kinds of distinctness or dimensionality of measurement: psychological and psychometric (Henning, 1992, p. 2). *Psychological* dimensionality pertains to the extent a psychological construct is unidimensional or multidimensional by nature. Psychological dimensionality is sample-independent and does not depend on the distribution of abilities in the people responding to the test items. Rather, it depends on the extent to which there is a match between test content and the construct theory (Henning, 1992).

On the other hand, *psychometric* dimensionality refers to the homogeneity of item score variances, which is independent of the psychological dimensionality. Psychometric unidimensionality can be present even when a test measures a variety of correlated underlying psychological dimensions, or it can be absent when the test is intended to measure a single construct. In this sense, psychometric dimensionality, like internal consistency reliability, is sample-dependent, influenced by measurement error, and fluctuates with changes in the distribution of abilities in the tested sample of test takers (Henning, 1992).

As mentioned before, in SLA the two constructs of L2 EXK and IMK are considered psychologically distinct, and the aim of validation studies is to investigate whether, for example, different kinds of GJTs can measure these two constructs in a psychometrically distinct way. Thus, throughout this paper, it is assumed that L2 EXK and IMK are psychologically distinct, and what needs to be verified is whether measures of different types can tap these two constructs in a psychometrically distinct way.

Factor-Analytic Validation Studies

A series of validation studies on different measures of L2 EXK and IMK using factor analysis was inspired by Ellis (2005). He designed several measures of EXK and IMK and used several quantitative methods to verify their construct validity. Ellis operationalized the two constructs following seven criteria: degree of awareness, time available, focus of attention, systematicity and certainty, meta-language, and learnability. Then, Ellis developed five "distinct" measures of IMK and EXK: an elicited imitation task (EI), an oral narrative task (ON), a timed grammaticality judgment tasks (TGJT), an untimed grammaticality judgment tasks (UGJT), and a metalinguistic knowledge test (MKT).

To complete the EI, the test takers first listened to a set of grammatical and ungrammatical sentences once (timed) and then indicated whether they agreed or disagreed with the propositions expressed in the sentences (drawing attention to meaning). Finally, the learners were asked to repeat the sentences and were audio-recorded. The scoring of EI was based on whether learners successfully repeated/corrected the target structure in each sentence. For completing the ON, the learners read a story twice before being asked to retell the story orally (drawing attention to meaning) in three minutes (timed). Their narratives were audio-recorded, and the scoring was based on the percentage of correctly supplied target structures. For the TGJT, the learners had to judge whether a sentence was grammatical or ungrammatical (drawing attention to form) within a fixed time limit (timed). For the UGJT, the learners judged the grammaticality of the sentences (drawing attention to form), but with no time limit (untimed). Finally, for the MKT, the test takers were presented with ungrammatical sentences and required to select the rule that best explained each error out of four choices provided (drawing attention to form) with no time limit (untimed).

As described above, using primarily two criteria, whether attention was drawn to form or meaning and/or whether there was a presence or absence of time pressure, Ellis (2005) hypothesized that the UGJT and MKT would tap EXK because they are both untimed and draw attention to form. On the other hand, the ON and EI would tap IMK because they are both timed and draw attention to meaning. TGJT, however, was hypothesized to tap IMK because it was timed.

Ellis (2005) subjected the test battery scores to a principal component analysis (PCA), which resulted in a two-component solution "confirming" his hypotheses. However, Ellis's study suffered from a few data analysis flaws. The most notable was that Ellis approached the data analysis with a set of hypotheses, making the use of a PCA inappropriate and the use of a CFA more appropriate instead. To rectify this drawback, Ellis and Loewen (2007) reanalyzed the data from Ellis (2005) first through a PCA and later also a CFA. In the CFA, they tested their hypothesized two-factor model with UGJT and MKT loading on the EXK factor, and the TGJT, ON, and EI loading on the IMK factor. Ellis and Loewen also tested an alternative model in which production and decision were considered as the two underlying factors, with the EI and ON loading on the former and the two GJTs and MKT loading on the latter. By finding an acceptable fit for the hypothesized model and rejecting the alternative model, Ellis and Loewen "confirmed" the findings in Ellis (2005).

However, although the CFA results revealed an acceptable two-factor solution, the correlation between the two factors was .51 and statistically significant. Bivariate correlation coefficient of an approximate value of .5 indicates a large effect size, representing evidence against discriminant validity (Swank & Mullen, 2017). Therefore, alternative models examining the psychometric distinctness of scores obtained from measures included in Ellis and Loewen (2005) were necessary. One theoretically and empirically plausible alternative model that should have been tested was a one-factor model with all measures loading on a single construct. Finding an acceptable fit for this one-factor model could have provided counter evidence for discriminate validity, which is required to claim that several testing instruments are distinct measures of distinct constructs. Testing this one-factor model is in lieu of testing an important rebuttal against the backing to the hypothesized structure obtained from the two-factor model. As said before, Ellis and Loewen (2007) did test one alternative model (decision vs. production); however, this alternative model was irrelevant to the purpose of their study, i.e., examining if the measures in the battery were distinct measures of EXK and IMK.

In addition, method effects, defined as a range of factors that can affect test performance and jeopardize test validity (Bachman & Savignon, 1990), were not accounted for in Ellis and Loewen (2007). When two-factor CFA models are confirmed to be the best fit to the data, the question would be whether the two separate factors were the result of differences in the methods of measurement of the tests loading on different factors or whether the tests tapped separate underlying constructs regardless of the differences in their methods of measurement. CFA enables researchers to test the method effects, for example, by specifying error covariations. This can account for the additional covariation among measured variables, not explained by the underlying construct resulting from, for example, similarities in measurement methods (Brown, 2006, p. 46).

The method effects that Ellis and Loewen (2007) could have considered is the similarity between the measurement methods of the MKT and UGJT. For one, unlike EI and ON, both of these tasks drew learners' attention to form rather than meaning. Also, unlike TGJT, which also drew attention to form, MKT and UGJT were both untimed. More importantly, for the UGJT, only ungrammatical scores were included in the analyses, making its results closely comparable to the ones from the MKT. In MKT, only ungrammatical sentences were presented to the learners, who were asked to choose from multiple options the rule that best explained the reason for the ungrammaticality of the sentences. Thereby, among the five measures, MKT and UGJT were most similar in their methods of measurement because both tasks asked learners to think about the ungrammaticality of sentences in an untimed condition. The only difference between the two was that, for the MKT, the reason for the ungrammaticality also needed to be chosen from a set of rules. Although EI and ON were both meaning-oriented and timed, their response types were very different: While for the former, the test takers were just required to repeat the isolated sentences verbatim, for the latter, the learners had to reconstruct a whole story in an extended production of language. However, the TGJT was different from both groups of measures. Unlike EI and ON, it drew attention to form, and unlike UGJT and MKT, it was timed. This means, a reasonable error covariation for accounting for the method effects would be the one between UGJT and MKT.

Another major validity study in this area, which replicated the results from Ellis (2005) and Ellis and Loewen (2007), was Bowles (2011). In terms of using CFA, this study also had the same methodological pitfalls as Ellis and Loewen. Bowles developed five tests of Spanish by following Ellis's (2005) criteria. A CFA was conducted to examine the factorial structure of the test battery. Bowles' results corroborated the two-factor model in Ellis and Loewen, but with a .89 correlation between the two factors. Although it was not reported whether the correlation was statically significant, its magnitude (an effect size) implies that the two factors were not psychometrically distinct. Thus, Bowles also should have tested a CFA alternative model in which all measures loaded on a single factor while accounting for the method effects.

Zhang (2015) was another replication study of Ellis and Loewen (2007) that used CFA to investigate the construct validity of a battery of measures of EXK and IMK that included a MKT, an UGJT, a TGJT, and an EI task. For developing the tests, Zhang also followed the Ellis's (2005) criteria and used the same target structures. Zhang tested two CFA models, one testing the two-factor IMK/EXK model reported in Ellis and Loewen, and the other testing a model in which the grammatical sentences of both TGJT and UGJT loaded on the IMK factor along with the EI task and the ungrammatical sentences of both types of GJTs loaded on the EXK factor along with the MKT. The latter model was inspired by Gutiérrez (2013), where it was hypothesized that this is the stimulus type (i.e., grammaticality of the GJT sentences) that determines whether GJTs measure IMK or EXK. According to Gutiérrez, GJT grammatical sentences and ungrammatical sentences tap IMK and EXK, respectively. The CFA in Zhang revealed that the two-factor model with the EI and TGJT loading on the IMK factor and the UGJT and MKT loading on the EXK factor was the best fit to the data. However, the correlation between the two factors in this model was .86 and statistically significant, which implies that the two factors were not psychometrically distinct. Similar to the previous studies, Zhang should have tested a rival one-factor model, which also accounted for the method effects.

In short, none of the above studies examined theoretically and empirically plausible alternative models in their CFAs. We agree with Isemonger (2007, p. 109), who warned that in CFA studies, "it is important that alternative models are tested because the fit of a particular model does not preclude the possibility that other untested models fit better." Even if a hypothesized model fits the data well, there might still be alternative or even equivalent models that fit the data equally well but with extremely different theoretical implications and interpretations (Hershberger & Marcoulides, 2013). Not testing for these alternative models can lead researchers to establish false causal relations and accept equally false implications.

In more recent years, several factor-analytic validation studies on the measures of IMK and EXK did actually test rival CFA models to challenge their own hypothesized structural relations in their test batteries. For example, Vafaee, Suzuki, and Kachinske (2017) tested 20 different CFA models to examine the construct validity of a battery of measures that included a MKT, an UGJT, a TGJT, a self-paced reading task (SPRT) and a word-monitoring task (WMT).

They had hypothesized that GJTs, regardless of their time conditions and/or stimulus type, are too coarse to be measures of IMK (because they still draw attention to form) and that TGJTs tap proceduralized EXK, at best. On the other hand, they had hypothesized that reaction-time measures like SPRTs and WMTs are more sensitive measures of IMK. The results of their CFA supported their hypotheses because the model that had the best fit to their data was a two-factor model with the MKT, UGJT, and TGJT loading on the EXK factor, and the SPRT and WMT loading on the IMK factor. In this model, the correlation between the two factors was .26 and statistically non-significant, which presented more convincing evidence that the two factors were measured in a psychometrically distinct way. The strength of Vafaee et al. (2017) was that they only retained their hypothesized model after testing 19 rival models including the one-factor ones in which the method effects were accounted for.

Although Vafaee et al. (2017) and other studies with the same level of methodological rigor (e.g., Suzuki, 2017) can be employed as successful models for the implementation of CFA in its full capacity, the focus of these studies was not the methodological issue that the current paper is trying to address. For a more focused demonstration of the validity consequences of not using CFA appropriately, the current study examines how validity conclusions about measures of EXK and IMK may lack empirical support if the researchers do not test rival CFA models to rebut their own hypothesized models.

THE PRESENT STUDY

Using the informal-argument structure, embedded in the argument-based validation framework (Kane, 2006), the current study was an attempt to show how CFA should be conducted in its full capacity and how validity conclusions about the nature of data collected by measurement instruments in Ellis and Loewen (2007), Bowles (2011), and Zhang (2015) were unwarranted because they did not test plausible CFA alternative models, or rebuttals, against their hypothesized models. These studies were selected from a number of factor analytic validation studies because they included the same set of measures and tested the same hypothesized models. The selected studies claimed that MKTs and UGJTs measure EXK, while TGJTs, ONs, and EIs measure IMK. These studies also provided backing for their claim by testing and retaining two-factor CFA models, implying that the measures were distinct measures of EXK and IMK. However, one important theoretically and empirically plausible alternative model rebutting this claim is a one-factor model that shows that such a psychometric distinction does not exist. By imposing one-factor alternative models on their data, the current study sought to answer the following research question: Do one-factor models fit the data as well as the two-factor models retained in the original studies?

It should be noted that the current study highlights the importance of testing CFA alternative models within Kane's (2006) argument-based validation framework. However, regardless of this framework, testing alternative models is a necessary step in conducting CFA, unless CFA is used for a "strictly conformity (SC)" situation (Jöreskog, 1993, p. 295). In this situation, the researcher proposes a single model and, after testing the model with empirical data, either accepts or rejects it. However, the strictly conformity situation is very rare because, in practice, few researchers accept or reject a model without suggesting an alternative. Thus, regardless of a particular validation framework, testing alternative models is an indispensable step in the CFA process.

Analysis

The covariance matrices reported in Ellis and Loewen (2007), Bowles (2011), and Zhang (2015) were used as data for the current analyses. For details about the instruments, participants, and data collection procedures, the readers are invited to refer to the original studies. For each of the above studies, the following CFA models were tested: (1) the hypothesized and retained two-factor model in the original study, (2) the alternative one-factor model with no error covariances, and (3) the alternative one-factor model with the appropriate error covariances specified a priori.[5]

The Robust Maximum Likelihood (RML) was used as the method of model parameter estimation, and a profile of model fit tests and indices recommended by Hu and Bentler (1999) and Mueller and Hancock (2008) was used to evaluate the models. Chi square (***χ2***), with its degrees of freedom and *p*-value, was checked. For a good model fit, the chi-square should not be statistically significant at a .05 level. However, in large samples and complex models, the chi-square is usually significant and not very informative. For this reason, the following descriptive fit indices were also used: the standardized root mean square residual (SRMR< .08), the root mean square error of approximation (RMSEA< .06), and the comparative fit index (CFI> .95). To compare models statistically, the chi-square difference (**Δ*χ2***) test was used. Mplus version 7 (Muthén & Muthén, 2012) was used to run the CFAs.

RESULTS

Ellis and Loewen (2007)

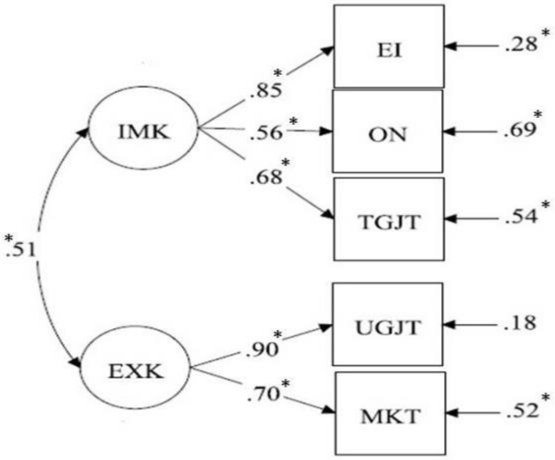
The model fit indices for this set of analyses are presented in Table 1. Model 1 (Figure 2) replicated the original two-factor model. In this model, MKT and UGJT loaded on the EXK factor, and EI, ON, and TGJT loaded on the IMK factor. In Figure 2, and the subsequent figures, the standardized model parameter estimates are presented, and the asterisks indicate the statistically significant parameter estimates.

TABLE 1
Summary of model fit indices for Ellis and Loewen (2007)

Index	CFI	RMSEA	SRMR	Chi-square	**Δ*χ2***
Criterion	≤ .95	≥ .06	≥ .08	None significant	
Model 1	1	0	.02	1.47 (4)	
Model 2	.79	.25	.1	32.5 (5)	31.03 (1), *p*<.05
Model 3	1	0	.02	1.47 (4)	0 (0)

[5] For all three studies, a rival two-factor model was also tested. In this model, the appropriate error covariances were added. However, for this model in Ellis and Loewen (2007) and Bowles (2011), the latent variable covariance matrix (PSI) was not positive definite because the correlation between the two factors was one. This means that adding the error covariances to the two-factor models actually turns them into single-factor models. For Zhang (2015), due to the limited available unique data points, this two-factor model with specified error covariance was under-identifiable (i.e., zero degree of freedom).

FIGURE 2
CFA Model 1 based on data from Ellis & Loewen (2007)

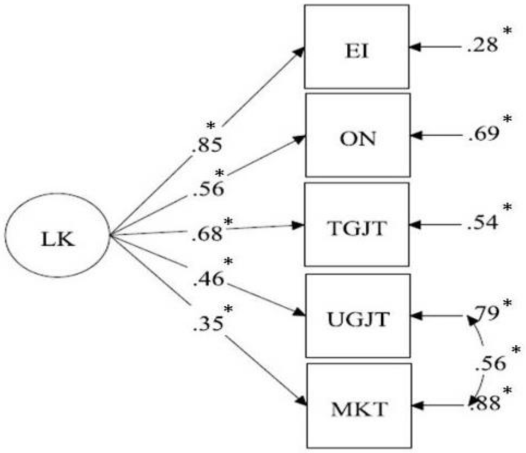


Model 1 fit the data well and all the model parameters were statistically significant, except for the error variance for the UGJT. However, this result does not rule out the plausibility of alternative models, especially because the correlation between the two factors was .51 and statistically significant. This motivated testing a one-factor alternative model.

First, we tested an alternative model with all measures loading on one factor (Model 2), labeled language knowledge (LK). We used the term LK to indicate that the measures did not tap distinct constructs of EXK and IMK. Rather, they were different measures of a single underlying factor that we tentatively called language knowledge.

However, this model got significantly worse in comparison to Model 1, as evidenced by (***Δχ2*** = 31.02, *df*= 1, *p*<.05). Thus, a second alternative one-factor model (Model 3) was specified by adding an error covariance between UGJT and MKT (Figure 3). This error covariance was added to account for the method effects. Both UGJT and MKT are untimed measures that draw attention to form. In both of these two tasks, the test takers have enough time to think explicitly about the grammaticality of the sentences.

FIGURE 3
CFA Model 3 based on data from Ellis & Loewen (2007)



Adding the error covariance resulted in an almost perfect fit for Model 3 with identical goodness-of-fit indices as in the retained two-factor model in Ellis and Loewen (2007) (i.e., Model 1 here) with the same χ^2 value, degrees of freedom, p values, and descriptive fit indices. This indicates the existence of an equivalent alternative model alongside their retained model (Kline, 2011; Raykov & Penev, 2001). Also, after adding the error covariance, the error variance for UGJT became statistically significant, yet another piece of evidence for the existence of method effects in the relationship between UGJT and MKT.

Bowles (2011)

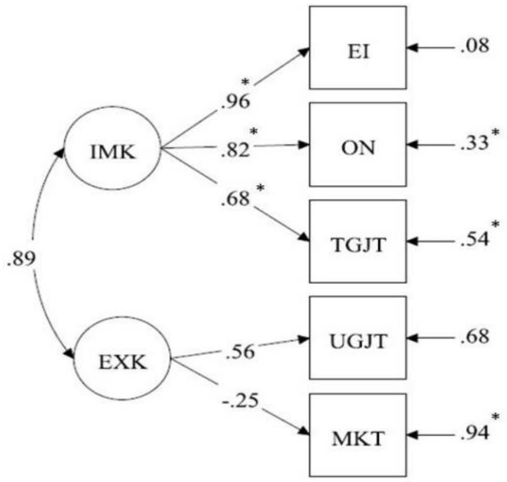
The model fit indices for this set of analyses is presented in Table 2. First, Model 1 (Bowles' (2011) original model) (Figure 4) was tested, with the MKT and UGJT loaded on the EXK factor, and the EI, ON, and TGJT loaded on the IMK factor.

TABLE 2
Summary of model fit indices for Bowles (2011)

Index	CFI	RMSEA	SRMR	Chi-square	**Δ*χ2***
Criterion	≤ .95	≥ .06	≥ .08	None significant	
Model 1	.98	.06	.07	4.31 (4)	
Model 2	1	0	.07	4.34 (5)	.03(1), *p*>.05
Model 3	.98	.06	.07	4.31 (4)	0 (0)

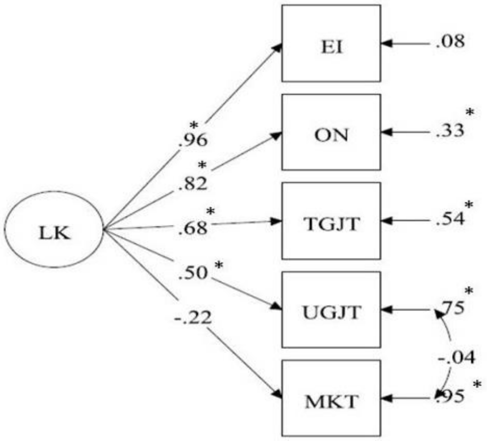
Although this model fit the data well, the loadings of the UGJT and MKT on their underlying factor, the error variances for EI and UGJT, and the correlation between the two factors were not statistically significant. In this model, the correlation between the two factors was .89. Although statically non-significant, which may be due to the extremely small sample size (i.e., *N*= 30), this magnitude of correlation implied the two factors may not be distinct. This motivated testing two one-factor alternative models.

FIGURE 4
CFA Model 1 based on data from Bowles (2011)



In Model 2, all measures loaded on one factor (i.e., LK), and in Model 3 (Figure 5), an error covariance between UGJT and MKT was added.

FIGURE 5
CFA Model 3 based on data from Bowles (2011)



As seen in Table 2, results from testing Models 2 and 3 revealed an acceptable fit for both of these alternative models. As evidenced by (**Δ*χ2*** = .03, *df*= 1, *p*>.05), the one-factor model with no error covariance (Model 2) was already equally plausible as the original two-factor model (Model 1).

However, the fit indices for the second alternative model with the error covariance (Model 3) revealed the presence of an equivalent model to Model 1. Additionally, in Model 3, although the error covariance between UGJT and MKT was not statistically significant, adding this covariance changed the loading of the UGJT on its underlying factor from statistically non-significant to significant. The loading of the MKT on the underlying factor and the error variance of EI, however, remained non-significant. Model 3 with one factor (more parsimonious) and larger number of statistically significant parameter estimates was considered superior to the two-factor Model 1.

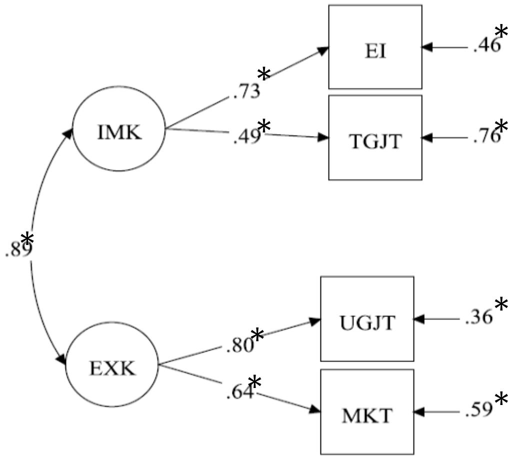
Zhang (2015)

The model fit indices for this set of analyses are presented in Table 3. Model 1 (Figure 6) corresponds to the model that was chosen as the best fitting model in Zhang (2015).

TABLE 3
Summary of model fit indices for Zhang (2015)

Index	CFI	RMSEA	SRMR	Chi-square	**Δ*χ2***
Criterion	≤ .95	≥ .06	≥ .08	None significant	
Model 1	1	0	0	.06 (1)	
Model 2	1	0	.02	.87 (2)	.81(1), *p*>.05
Model 3	1	0	0	.061 (1)	0 (0)

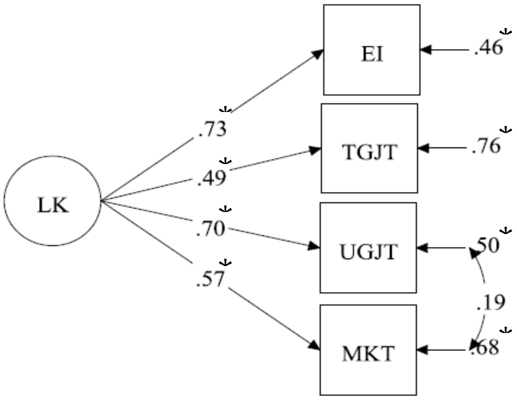
FIGURE 6
CFA Model 1 based on data from Zhang (2015)



In this model, the MKT and UGJT loaded on the EXK factor, while the EI and TGJT loaded on the IMK factor. This model fit the data well, and all the parameter estimates were statistically significant. However, the possibility of the existence of plausible one-factor alternative models was not ruled out because the correlation between the two factors was .89 and statistically significant. This motivated testing several one-factor alternative models.

In Model 2, all the measures loaded on one factor (i.e., LK), and in Model 3 (Figure 7), an error covariance between UGJT and MKT was added. The fit of Model 2 became slightly worse than Model 1, but the difference between the fit of the two models was not statistically significant (**Δ*χ2*** = .81, *df*= 1, *p*>.05). This finding already shows that the tests in this battery did not tap two distinct constructs.

FIGURE 7
CFA Model 3 based on data from Zhang (2015)



Even more convincing, as seen in Table 3, Model 3 had identical fit indices as Model 1, which is strong evidence against the discriminate validity claim proposed in Zhang (2015). In model 3, all the parameter estimates were statistically significant, except for the error covariance between UGJT and MKT. Although statistically non-significant (maybe due to an insufficient sample size), the addition of the error covariance between UGJT and MKT turned model 3 into an equivalent model of Model 1.

DISCUSSION

The current study intended to raise awareness among SLA researchers about the benefits of adopting a conceptual validation framework to make informed decisions about, among other things, conducting CFA in its full capacity. Validations studies on measures of EXK and IMK such as Ellis and Loewen (2007), Bowles (2011), and Zhang (2015) examined whether, for example, manipulating GJTs time conditions can turn them into distinct measures of EXK and IMK. By finding an acceptable fit for the two-factor models, these researchers provided support for their own hypotheses. The studies concluded that: (1) MKTs and UGJTs are measures of EXK because when completing them, test takers have enough time to pay attention to form and tap their EXK to decide on the grammaticality of sentences, and (2) TGJTs are measures of IMK because when completing them, test takers do not have enough time to use their EXK, so they rely on their intuition or IMK. For ONs and EIs, the two criteria that make them measures of IMK are the presence of time pressure and the drawing of attention to meaning rather than form. However, none of these researchers tested alternative models against their own hypothesized models. Ignoring alternative models is a form of confirmation bias, whereby, when only a single model is tested, premature confirmation is given to the model and no other explanation of the data is offered (Shah & Goldstein, 2006). In case of Ellis and Loewen (2007), Bowles (2011), and Zhang (2015), testing one-factor models (i.e., testing rebuttals against their validity conclusions) was a plausible alternative because the magnitude of the correlations between the two factors retained in these studies indicated the lack of distinctness between the two factors. Because these researchers did not test these one-factor models, we sought to examine the validity of their conclusions by answering the following research question: Do one-factor CFA models fit the data as well as the two-factor models retained in the original studies?

In our analyses, we also added the error covariances between measured variables with similar testing methods (i.e., time condition and drawing attention to form or meaning) in the one-factor models. This was deemed important because in CFA adding the error covariances accounts for the covariation between variables resulted from method effects (Brown, 2006, p. 46).

Our results demonstrated that not only were there no statistically significant differences between the two-factor models retained in the original studies and our one-factor alternative models with error covariances, but also that these alternative models were equivalent to the originally retained models. The existence of statistically equivalent models is highly problematic for inferring structural relations and substantive interpretations from the CFA models (Hershberger & Marcoulides, 2013). Equivalent models, as opposed to alternative models that are not statistically different, pose a special threat to structural inferences because “any of them can never be supported without all of them being supported” (Hershberger & Marcoulides, 2013, p. 8).

What our analyses demonstrated is that in the case of Ellis and Loewen (2007), Bowles (2011), and Zhang (2015), the existence of equivalent models has brought into question the studies' conclusions. It can be said that these studies did not provide compelling evidence demonstrating that measures included in their test batteries were distinct measures of EXK and IMK.

CONCLUSIONS

Validity is not an inherent quality of a measurement tool itself. Rather, it is an argument for or against the plausibility and appropriateness of interpretations and uses made based on test scores drawn on empirical data and logical reasoning. Therefore, the aim of validation is to provide an overall evaluation of the plausibility and appropriateness of these interpretations and uses. This can be achieved through a coherent analysis of all of the empirical evidence and logical reasoning for and against the plausibility and appropriateness of the proposed interpretations and uses (Cronbach, 1988).

This means that no one can validate a measure that fits all purposes. Instead, for every unique interpretation and use of test scores, a unique validity argument should be built. For conducting validation studies and constructing validity arguments, Kane (2006) proposed an argument-based validation framework. This framework can be used as a guideline on how to collect a variety of evidence to develop a validity argument that justifies the plausibility and appropriateness of the interpretations and uses of test scores. The minimum benefit of adopting this framework is being aware that it is not sufficient merely to collect supporting evidence for the validity claims. Collecting evidence against these claims, i.e., testing rebuttals, is also a necessary step in the validation process.

The aim of the current paper was not to build a validity argument for or against the plausibility and appropriateness of interpretations and uses made based on scores or data from any particular measure of L2 EXK or IMK. As explained before, any validity argument should be pertinent to a specific interpretation based on data collected for a specific purpose and is limited to the context of a specific study. For example, the resultant data from using a measure, for example a GJT, from native speakers of a language and the interpretations based on these data should be limited to this particular population. The same is true if the same instrument is used to collect data from a different population, for example, the learners of a foreign language.

Nonetheless, SLA researchers can still run validation studies on different kinds of measures of L2 EXK and IMK with different populations to gain a general understanding of the nature of data collected by these measures. However, building a unique validity argument for or against the plausibility and appropriateness of interpretations and uses based on a specific measure in a specific context is a burden on the shoulders of the test score users. Every validity argument is unique in its own right.

REFERENCES

American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Lawrence Erlbaum Associates Publishers.

Bachman, L. F., & Savignon, S. J. (1990). *Fundamental considerations in language testing* (Vol. 107). Oxford University Press.

Bowles, M.A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, *33*(2), 247–271.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. *Test Validity*, 3–17.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.

Chapelle, C. A. (2013). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). Routledge.

DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 119–138). Wiley-Blackwell.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*(2), 305–352.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, *27*(2), 141–172.

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*(1), 119–126. https://doi.org/10.1017/S0272263107070052

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 35*(3), 423–449. https://doi.org/10.1017/S0272263113000041

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1–11.

Hershberger, S. L., & Marcoulides, G. A. (2013). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds), *Structural equation modeling: A second course* (pp. 3–41). Information Age Publishing Inc.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Isemonger, I. M. (2007). Operational definitions of explicit and implicit knowledge: Response to R. Ellis (2005) and some recommendations for future research in this area. *Studies in Second Language Acquisition*, *29*(01), 101–118.

Jöreskog, K., G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Sage.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). The Guilford Press.

Krashen, S. (1994). The input hypothesis and its rivals. In N. Ellis (Ed.), *Implicit and explicit learning of language* (pp. 45–77). Academic Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62.

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Sage Publications, Inc.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus statistical modeling software: Release 7.0.* Muthén & Muthén.

Raykov, T., & Penev, S. (2001). The problem of equivalent structural equation models: An individual residual perspective. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 297–321). Lawrence Erlbaum.

Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, *24*(2), 148–169.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). IAP Information Age Publishing.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.

Swank, J. M., & Mullen, P. R. (2017). Evaluating evidence for conceptually related constructs using bivariate correlations. *Measurement and Evaluation in Counseling and Development*, *50*(4), 270–274.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*(1), 59–95.

Zhang, R. (2015). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, *37*(3), 457–486.