Educational and Psychological Measurement
70(2) 323–339

DOI: 10.1177/0013164409344507
http://epm.sagepub.com



# Attempted Validation of the Scores of the VARK: Learning Styles Inventory With Multitrait–Multimethod Confirmatory Factor Analysis Models

**Walter L. Leite,[1] Marilla Svinicki,[2] and Yuying Shi[1]**

**Abstract**

The authors examined the dimensionality of the VARK learning styles inventory. The VARK measures four perceptual preferences: visual (V), aural (A), read/write (R), and kinesthetic (K). VARK questions can be viewed as testlets because respondents can select multiple items within a question. The correlations between items within testlets are a type of method effect. Four multitrait–multimethod confirmatory factor analysis models were compared to evaluate the dimensionality of the VARK. The correlated trait–correlated method model had the best fit to the VARK scores. The estimated reliability coefficients were adequate. The study found preliminary support for the validity of the VARK scores. Potential problems related to item wording and the scale's scoring algorithm were identified, and cautions with respect to using the VARK with research were raised.

**Keywords**

VARK, learning styles, multitrait–multimethod confirmatory factor analysis, method effect, testlet

Teachers at all levels of education recognize that students do not all respond equally well to the same instructional methods, and therefore, teachers are intrigued by the possibility of providing instruction matched to those learner differences. Learners,

[1]University of Florida, Gainesville, FL, USA
[2]University of Texas, Austin, TX, USA

**Corresponding Author:**
Walter L. Leite, Department of Educational Psychology, University of Florida, 1215 Norman Hall, Gainesville, FL 32611, USA

too, have experienced in themselves that apparent differential preference for one method of learning over another. As a result of these almost universal observations about learning, researchers in education have proposed and evaluated many theories and instruments intended to help either the teacher or the learner become aware of learning preferences so that the instructional environment can be tailored to learner needs. Jonassen and Grabowski (1993) provided a fairly detailed annotated listing of instruments intended to inform researchers, educators, and learners alike about the range from which they could choose. Sternberg and Zhang (2001) have published a very comprehensive edited volume discussing the history, traditions, research, and theory around the topic of thinking, learning, and cognitive styles. Both these books represent important points along the path of the phenomenon of learning styles.

One of the more popular instruments of this type is the VARK, developed by Fleming (2001; Fleming & Mills, 1992). Its popularity comes from its face validity, its simplicity, its ease of use, and the wealth of learning materials that have been designed to accompany it. Most users have very practical reasons for using it. Many want to increase awareness and conversation about learner differences as a precursor to encouraging teachers to use more varied instructional methods. Some want to help students become aware of their own preferences so that those students can better plan their own learning strategies to take advantage of their strengths. For these uses, the instrument is excellent. However, many individuals are interested in using the VARK as an input or outcome measure in research, and it is these uses that require a more rigorous evaluation of its psychometric properties. This is the purpose of the analyses reported in this paper. We begin by reviewing the literature on learning styles in general and the VARK in particular and then provide the results of analyses done on the latest version of the instrument. Finally, we offer advice and caution about future uses of the instrument.

## *Learning Styles*

The term *learning styles* was first cited in the literature more than 50 years ago (Thelen, 1954). Despite the fact that this area of study is a half-century old, there is still little agreement about a precise definition of learning styles (Anderson & Adams, 1992). Furthermore, researchers also disagree about the nature of the relationship and the overlap between the concepts of learning styles, cognitive styles, and learning ability. Several overviews of the various conceptualizations of learning style have been published, which have helped sort out the field somewhat (Jonassen & Grabowski, 1993; Sternberg & Zhang, 2001). Most modern style theories focus more on the cognitive process aspects of learning style, such as those that take a wholistic/analytical approach to learning on one dimension and a visual/verbal approach to learning on another dimension (Riding, 2001). This dual process model of cognitive style seems to incorporate numerous dimensions from other cognitive style models that use the same concepts but different terms to refer to the dimensions.

A major disagreement in the literature is about what type of learning preferences should be included under the umbrella of learning styles. Some researchers see learning styles as referring only to perceptual preferences, whereas others view them

as capturing any preference that affects learning. For example, Keefe (1987) viewed learning style as the broader term that includes cognitive, affective, and physiological styles. Dunn, Beaudry, and Klavas (1989) argued that learning styles are biologically and developmentally defined characteristics that include perceptual preferences, instructional environment preferences, sociological preferences, mobility needs, and time-of-day preferences. Kolb (1976) proposed a more strict view of learning styles, considering them to be the individual's preferred way to perceive and process information. Kolb's (1976) Experiential Learning Model conceptualizes learning styles as individual preferences with respect to four modes: concrete experience, reflective observation, abstract conceptualization, and active experimentation (Henson & Hwang, 2002).

The disagreement on the definition of learning styles has resulted in a body of research that is very fragmented, using different instruments to measure different constructs under the heading of learning styles (Sternberg & Zhang, 2001). One thing that most learning style instruments have in common is a lack of solid research on their psychometric properties. Learning style instruments tend to be constructed in isolation from one another without much attempt to validate their underlying constructs, but because the concept of style appeals so strongly to educators and learners alike, there is often a rush to implementation without adequate analysis of the properties of an instrument.

## *Psychometric Issues in Learning Style Instruments*

The fact that the nature of the latent constructs measured by the existing learning styles instruments is still unclear makes it difficult to establish the validity of the scales' scores for their intended interpretations. Arguments about the validity of the scores of a learning styles instrument should be supported by multiple sources of evidence, such as test content, response processes, internal structure, relationships with other variables, and consequences of testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). An extensive collection of validity information for the scores of a learning style instrument would require several studies with both qualitative and quantitative analyses.

Among the possible sources of validity data, obtaining information with respect to internal structure usually requires psychometric analyses of large amounts of item-level data. Psychometric analyses may include evaluations of the dimensionality of the items, the items' difficulty parameters or thresholds, the items' discrimination parameters or factor loadings (Kamata & Bauer, 2008), measurement invariance (Reise, Widaman, & Pugh, 1993), and reliability (Raykov, 2001b).

The purpose of this article is to subject the VARK learning styles inventory (Fleming, 2001) to psychometric analyses aiming to collect evidence of validity with respect to internal structure. More specifically, we focus on evaluating the dimensionality and reliability of the VARK scores. The VARK represents an intuitively appealing perceptual preference dimension, usually conceived of as having three or more components, such as visual, auditory, and kinesthetic dimensions.

## *Perceptually Based Learning Style Instruments*

There are several learning style instruments that use perceptual preference as one of their factors (Cassidy, 2001). One instrument that has been extensively studied is Dunn, Dunn, and Price's (1985) Learning Styles Inventory. This instrument has subscales that are intended to measure categories such as environmental effects (sound, light, temperature), emotional characteristics, psychological characteristics, and sociological characteristics, such as preference for working with peers or alone. Important for this article is the final category of physical preferences, which includes the perceptual dimensions of visual versus auditory input. The research community that uses this scale is very large, and there have been studies investigating every component of Dunn et al.'s (1985) model of learning styles. Many studies showed statistically significant differences in achievement between students whose instruction was matched to their learning styles and students whose instruction was not, a key argument in favor of learning styles use (Dunn, Griggs, Olson, & Beasly, 1995). However, there were also studies that showed no statistically significant differences between these two conditions (Jonassen & Grabowski, 1993). Several studies showed trends among students of the same age or grade, but there was a large within-group variability in many of these studies (Dunn, Beaudry, & Klavas, 1989).

Fleming (2001; Fleming & Mills 1992) attempted to establish perceptual modes (e.g., visual, auditory, kinesthetic) as a measurable construct and is the focus of this article. He was influenced by research in neurolinguistic programming, suggesting that individuals receive information through sensory modalities and have sensory modality preferences. Fleming (2001) developed the VARK learning styles inventory, which aims to measure instructional preferences independent of personality characteristics, information processing strategies, and social interaction strategies in the classroom.

Fleming (2001) designed the VARK to measure four different perceptual preferences for the input of information, which are visual (V), aural (A), read/write (R), and kinesthetic (K). Visual individuals prefer to learn information presented in charts, graphs, and other symbolic devices instead of words. (Interestingly, the author defined the visual preference as not including pictures, movies, and videos because he argued that these presentations also involve kinesthetic, read/write, and aural perceptual modes.) Aural individuals prefer to learn from spoken lessons and talking, read/write individuals prefer to learn from printed text, and kinesthetic individuals prefer to learn through direct practice, which may also involve the other perceptual modes.

Despite its popularity, there has not been any rigorous attempt to establish the validity of the scores of the VARK questionnaire. Fleming (2001) mentioned the results of a principal components analysis in the VARK manual, where he extracted three components from four variables (i.e., visual, aural, read/write, and kinesthetic). However, this analysis examined the VARK preferences as composite variables and not the items of the questionnaire. Furthermore, obtaining three components from four variables does not accomplish a useful reduction of the data. Fleming did not report any estimate of the reliability of the VARK scores.

Because the VARK has been extensively used for advising students about their learning styles, its scores need to be more adequately validated. Our research questions were as follows:

1. Does the four-factor hypothesized structure of the VARK scale adequately explain the relationships between the observed scores on the VARK items? Answering this question required the evaluation of the fit of measurement models to VARK data.
2. Can adequate reliability estimates be obtained for the VARK scores? Estimation of reliability was performed under the framework of confirmatory factor analysis (CFA).

## Method

### *Sample*

For this analysis, the data collected with the VARK Web site during January 2007 were used. To control for possible cultural differences in the VARK, only students from the United States were included in the sample. Furthermore, only individuals who were taking the VARK for the first time were included in the sample. The sample size was 15,136. From this sample, 66.1% were female, 32.4% were male, and 1.5% did not declare their sex. With respect to age, 67.9% of the sample was 25 or younger, 26.3% was between 26 and 44 years old, 15.7% was 45 years or older, and 0.2% did not report age. Also, 43.9% was in a 4-year college or university, 41% was in a 2-year college, 9.1% was in high school, 5.7% chose the "other" category, and 0.3% did not respond. From the total sample, 925 (6.1%) did not respond to at least one question. These respondents were excluded from the analysis, resulting in a final sample of 14,211 individuals.

### *Instrument*

Data were collected with the online 16-question VARK Questionnaire, version 7 (Fleming, 2008). The instruction presented above the first question states, "Choose the answer which best explains your preference and circle the letter(s) next to it. Please circle more than one if a single answer does not match your perception." The VARK contains 16 questions with four response options. Each option is associated with a different learning style preference. An example question is shown below:

Do you prefer a teacher or a presenter who uses:
- Demonstrations, models or practical sessions
- Handouts, books, or readings
- Question and answer, talk, group discussion, or guest speakers
- Diagrams, charts, or graphs.

## Analysis

Because respondents to the VARK have the choice of selecting each option within a question, the VARK should be viewed as a questionnaire composed of 16 testlets of 4 dichotomous items each. A testlet is a group of items that is treated as a measurement unit in test construction, administration, and/or scoring. Testlets may be formed by grouping together items that share common stimuli (e.g., passages, pictures), format characteristics (e.g., multiple-choice or constructed response), stem, rubric, or rater (Lee, Brennan, & Frisbie, 2000). In the VARK, the items within each testlet share a common stem. Although uncommon in personality and attitudes assessment, testlets are frequently used in achievement tests (e.g., TOEFL, Iowa Test of Basic Skills, MCAT).

When CFA or item response theory (IRT) is used to analyze responses to items, the assumption of local independence is usually made. This assumption states that items are uncorrelated once their underlying latent constructs are accounted for. However, because items within a testlet share common stimuli, stem, format, rubric, or rater, their correlations cannot be accounted solely by underlying latent traits. Ignoring the lack of local independence of items within testlets tends to result in poor fit of the hypothesized model.

Scores from achievement scales based on testlets are usually analyzed with IRT models. To obtain local independence between items of a testlet, one possible solution is to consider each testlet as a single categorical item and fit a polytomous IRT model. Treating each testlet as a single categorical item is only possible for unidimensional testlets, which are testlets containing items measuring a single latent trait. Other IRT models have been proposed for testlets that assume unidimensionality (e.g., Wainer & Wang, 2000), but multidimensional models have also been proposed (e.g., Wang & Wilson, 2005).

The correlations between items caused by their grouping in the same testlet are a type of method effect. MTMM-CFA has been frequently used to model method effects. One advantage of using MTMM-CFA models with testlets is that they allow multidimensionality within testlets. This is the case of the VARK, where each testlet contains items measuring four latent traits. Furthermore, differently than IRT models, the relative appropriateness of several CFA-based measurement models can be easily compared (Lee, Dunbar, & Frisbie, 2001). The first MTMM-CFA model included latent factors for both traits and factors (Widaman, 1985). This model is known as correlated trait–correlated method (CTCM) model (see Figure 1). Tomás, Hontangas, and Oliver (2000) identified the CTCM model as the most popular approach to analyzing MTMM data. In this model, traits and methods are allowed to correlate among each other, but traits are not allowed to correlate with methods. The CTCM model has the advantage of allowing the total variance of the items to be decomposed into variance components due to traits, methods, and error. However, CTCM models may be difficult to estimate because of the frequent occurrence of underidentification, empirical underidentification, nonconvergence, and improper solutions (Lance, Noble, & Scullen, 2002). Furthermore, it is possible that method factors actually reflected trait variance in addition to method variance. A special case of the CTCM model, the correlated trait–uncorrelated method (CTUM) model performs better than the CTCM model with respect to identification, convergence, and solution propriety.

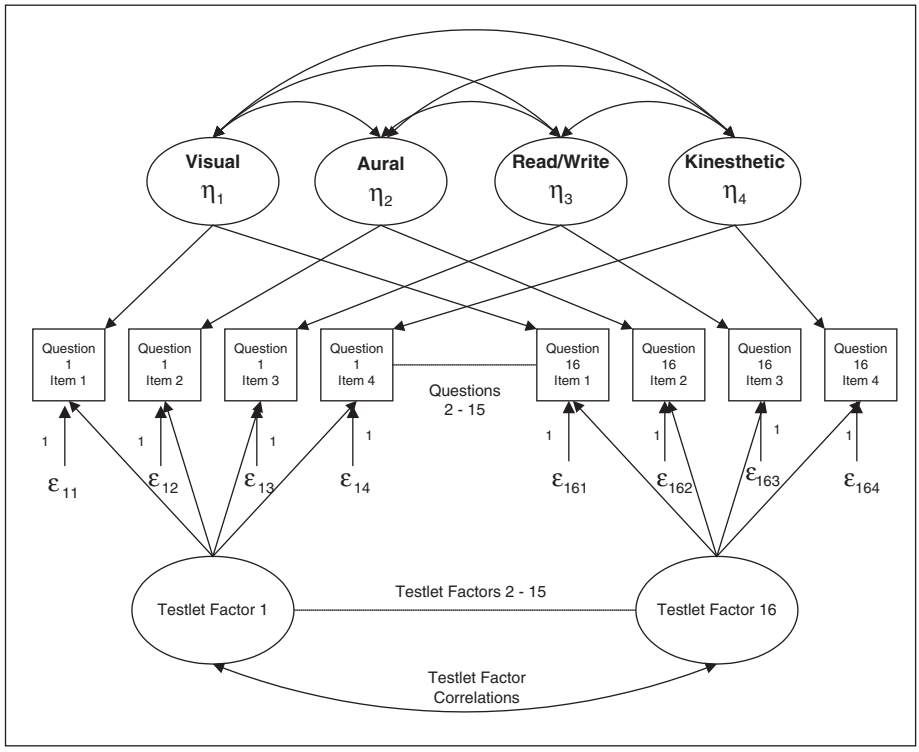


**Figure 1.** Correlated trait–correlated method model

Marsh (1989) proposed the correlated trait–correlated uniqueness (CTCU) model, where method effects are represented as correlated errors (see Figure 2). The CTCU model has the advantage of producing improper solutions less frequently than the CTCM model. However, the CTCU model has some shortcomings as compared with the CTCM model. First, the method effects are confounded with other sources of systematic error within a testlet, which does not permit the evaluation of the magnitude of method effects. Furthermore, correlations between methods are not allowed, which may result in poor model fit. Also, in the CTCM model a researcher may include covariates in the model to predict method effects, but in the CTCU model this is not possible.

Tomás et al. (2000) conducted simulation studies comparing the CTCM and the CTCU models in conditions with either one indicator or multiple indicators per trait–method combination. They found that with one indicator per trait-method combination the CTCU model obtained 100% convergent and proper solutions. Surprisingly, the CTCM model only resulted in nonconvergence with 4.81% of data sets. However, the CTCM model produced 48.88% of improper solutions. In conditions with one indicator per trait–method combination and a nonzero correlation between methods,

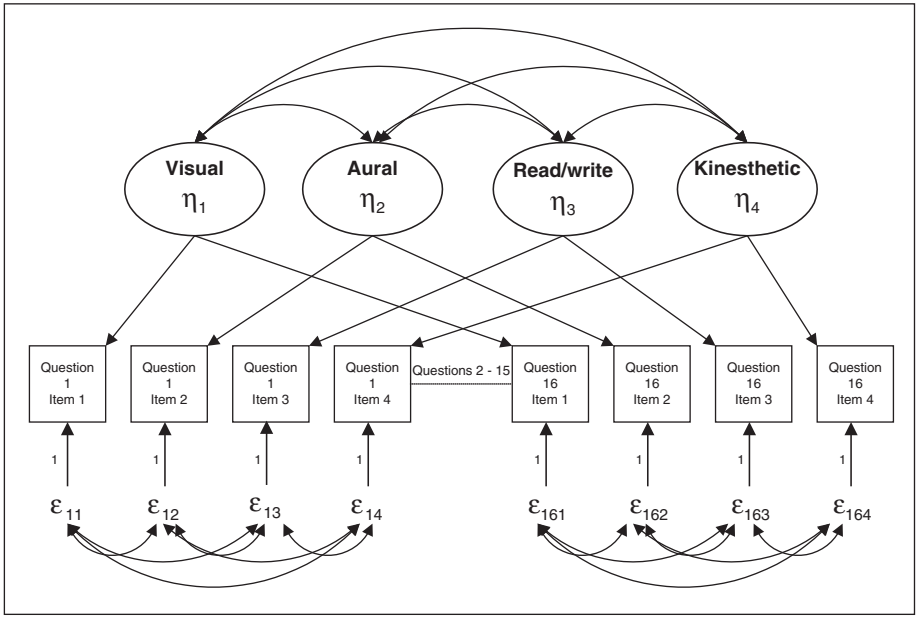


**Figure 2.** Correlated trait–correlated uniqueness model

the authors found that the CTCU model produces more accurate estimates of correlations between traits and factor loadings than the CTCM model. If the correlations between methods are zero, the CTCU and CTCM models were found to perform similarly. With multiple indicators per trait–method combination, the CTCM resulted in 3.5% of nonconvergence and 23.5% of improper solutions, while the CTCU model was found to be underidentified. Therefore, the authors recommended using the CTCU model in situations where there is a single indicator per trait–method combination and the CTCM model in situations where there are multiple indicators per trait–method combination.

The use of MTMMA-CFA models with testlets has been evaluated by Lee, Dunbar, and Frisbie (2001). They compared the performance of eight models with three different subscales of the Iowa Test of Basic Skills and a simulated data set. The CFA models evaluated included models that assumed essentially tau-equivalent measurements (i.e., equal factor loadings) and models that allowed congeneric items (i.e., factor loadings are free to vary). The authors found that congeneric models always outperformed essentially tau-equivalent models in terms of model fit. From the congeneric models evaluated, a model with correlated measurement errors (i.e., the CTCU model), a model with testlet factors (i.e., the CTCM model), and a model for testlets treated as categorical items performed best.

Eid (2000) proposed the correlated traits–correlated methods minus one (CT-C (M-1)) model for MTMM data (see Figure 3). In this model, the number of method factors is specified to be $m - 1$, where $m$ is the number of methods originally in a design. The unique part of his model is to define a method $k$ as a reference for all other methods. Suppose $T_{il}$ is the true score of trait $i$ measured by method $l$ and $T_{ik}$ is the true score of trait $i$ measured by method $k$. The method factor is defined as the part of $T_{il}$ that could not be predicted by $T_{ik}$. Here, the method $k$ was taken as a comparison standard. The author defined the method effect as

$$M_{il} = T_{il} - \mathrm{E}(T_{il}|T_{ik}), \tag{1}$$

where $M_{il}$ represented the effect of method $l$ with respect to a method $k$ and $\mathrm{E}(T_{il}|T_{ik})$ represents the part of $T_{il}$ that can be predicted by $T_{ik.}$.

The method first presented by Eid (2000) could only be applied to situations with one indicator per trait–method combination. Later, Eid, Lischetzke, Nussbeck, and Trierweiler (2003) expanded the CT-C(M-1) to account for multiple indicators per trait–method combination. The CT-C(M-1) model assumes that trait factors are uncorrelated with method factors and errors and that method factors are uncorrelated with errors. The CT-C(M-1) model has the advantage of dividing the item variance into variance due to the trait, method, and error, with the exception of the items of the reference method. However, the CT-C(M-1) model also has some limitations. First, the interpretation of the model parameters depends on the choice of reference method, which may be arbitrary. Also, model fit depends on the choice of the reference method. To address the model fit problem, Eid (2000) suggested fitting the model as many times as there are methods, switching the reference method at each run. The CT-C(M-1) model is the most recently proposed MTMM-CFA model and there have been no simulation studies comparing its performance against the CTCM and CTCU models.

We evaluated the dimensionality of the VARK using the CTCU, CTCM, CTUM, and CT-C(M-1) models. Because each of these models for MTMM data has some advantages and disadvantages and no model has emerged as clearly superior to the others, a comparison of models is justified. For all models, congeneric items were allowed. A model treating testlets as categorical items was not used because it does not allow for multidimensional testlets. Given that there was no theoretical reason for choosing any of the VARK testlets as a reference method for the CT-C(M-1) model, we used the first question as the reference method.

The software Mplus 4.1 (Muthén & Muthén, 2006) was used to fit the MTMM-CFA models. CFA is commonly performed using maximum likelihood (ML) estimation. However, ML estimation assumes that the items are continuous and normally distributed. Because the items of the VARK scale are dichotomous, we used the mean- and variance-adjusted weighted least squares estimator (WLSMV), which is the default robust estimator of the Mplus software for analyzing categorical indicators.

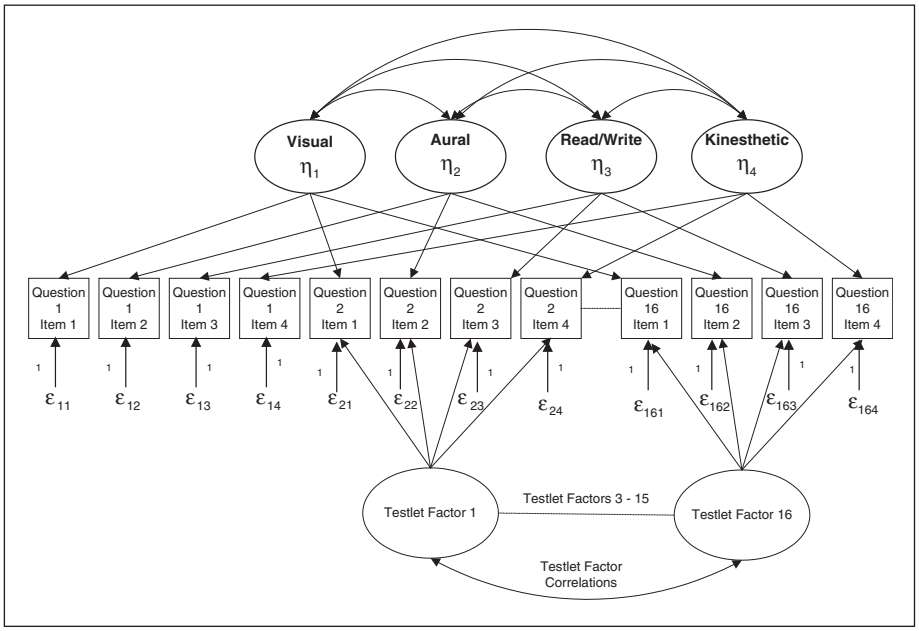


**Figure 3.** Correlated trait–correlated method minus 1 model

## Results

The MTMM-CFA models were fit to the data and their model fit was compared. The first model fit to the data was the CTCM model shown in Figure 1. This model turned out to be nonidentified. To solve the identification problem, we attempted fixing the correlations between methods at zero (the CTUM model). However, this model was also not identified. Therefore, a modified version of the CTCM model was evaluated. In this modified CTCM, one factor loading for each method factor was fixed to one, in addition to the factor variances which were fixed to one to set the scale of the latent variables. This modified CTCM model was identified and the solution was proper. The evaluation of the model fit resulted in $\chi^2(1090) = 34322.667$, $p < 0.01$, comparative fit index (CFI) = 0.708, Tucker–Lewis index (TLI) = 0.818, root mean square error of approximation (RMSEA) = 0.047, and standardized root mean square residual (SRMR) = 0.064.

Fitting the CTCU model displayed in Figure 2 to the scores of the VARK resulted in a nonpositive definite residual covariance matrix. The source of the problem was a perfect linear dependency between some errors of items of the kinesthetic scale. To solve this problem, the model was revised by fixing the error correlations that Mplus indicated as being problematic to the tetrachoric correlations between the items. From the 96 error correlations of items within testlets, 24 were fixed. The assessment

of model fit for the revised model with correlated measurement errors produced $\chi^2(1100) = 18779.444$, $p < 0.01$, CFI = 0.845, TLI = 0.905, RMSEA = 0.034, and SRMR = 0.049.

The CT-C(M-1) model shown in Figure 3 was also fit to the data, using Question 1 as the reference method. However, the model was not identified. The identification problem was solved by fixing one factor loading for each method factor to one, as well as the factor variances of the method factors. For the revised CT-C(M-1) model, the evaluation of model fit produced $\chi^2(1038) = 42634.880$, $p < 0.01$, CFI = 0.559, TLI = 0.735, RMSEA = 0.053, and SRMR = 0.074.

Comparing the model fit information for the MTMM-CFA models lead to the conclusion that the CTCU model had the best fit to the data. Therefore, in the discussion section we address only this model. The RMSEA, SRMR, and TLI, but not the CFI, supported the fit of the CTCU model. Given that agreement about fit from many fit indices is difficult to obtain with large multifactor models (Marsh, Hau, & Grayson, 2005), we concluded that there is preliminary evidence of adequate model fit for the CTCU model and provide more details about the parameter estimates for this model.

To demonstrate the consequences of inappropriately modeling scores of testlets by failing to account for method effects, we fit a CFA model with a diagonal residual covariance matrix. As expected, the model presented poor fit, $\chi^2(1119) = 51325.085$, $p < 0.001$, CFI = 0.634, TLI = 0.763, RMSEA = 0.056, and SRMR = 0.079. Because this model is nested within the CTCU model, we performed a chi-square difference test to evaluate whether the improvement of model fit with the CTCU model is statistically significant. Because these models were estimated with the WLSMV estimator of Mplus, the resulting chi-square statistic cannot be used for chi-square difference tests (Muthén & Muthén, 1998-2007). Therefore, the derivatives of the model with cross-loadings were used to obtain the correct chi-square difference test. The result, $\chi^2(68) = 60433.20$, $p < 0.001$, indicated that the improvement in fit with the CTCU model as compared with the CFA model ignoring method effects was statistically significant.

Using the CTCU model, the standardized loadings of the items on the VARK factors ranged from .24 to .76, with mean loadings of .51, .47, .50, and .41 for the visual, aural, read/write, and kinesthetic factors, respectively. The standardized loadings can be interpreted as correlations between the items and each VARK factor and are mostly of moderate size. The loadings for all items of the VARK are shown in Table 1. We also obtained the correlations between the VARK factors, which were moderate to strong in magnitude. These correlations are shown at the bottom of Table 1.

The parameter estimates from the CTCU model allowed us to estimate the reliability of the scores of VARK subscales. Cronbach's alpha was not used because it has been shown that it underestimates the reliability of the scores when the items of a scale are congeneric and there are correlated measurement errors (Komaroff, 1997; Raykov, 2001a). We obtained the reliability of the scores of VARK subscales with the following equation (Raykov, 2001b):

**Table 1.** Standardized Factor Loadings and Factor Correlations for the VARK

| Factor Loadings | Visual | Aural | Read/Write | Kinesthetic |
|---|---|---|---|---|
| Q1 | .415 | .468 | .401 | .325 |
| Q2 | .315 | .456 | .409 | .260 |
| Q3 | .522 | .374 | .475 | .559 |
| Q4 | .424 | .557 | .460 | .236 |
| Q5 | .760 | .424 | .518 | .258 |
| Q6 | .439 | .427 | .532 | .495 |
| Q7 | .590 | .541 | .609 | .415 |
| Q8 | .493 | .516 | .526 | .381 |
| Q9 | .579 | .499 | .531 | .442 |
| Q10 | .565 | .492 | .573 | .468 |
| Q11 | .423 | .425 | .424 | .320 |
| Q12 | .497 | .486 | .585 | .522 |
| Q13 | .656 | .497 | .631 | .518 |
| Q14 | .490 | .383 | .499 | .444 |
| Q15 | .497 | .544 | .470 | .452 |
| Q16 | .570 | .428 | .363 | .490 |
| Factor correlations | | | | |
| Visual | — | .601 | .463 | .799 |
| Aural | — | — | .437 | .737 |
| Read/write | — | — | — | .330 |

$$\rho_s = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2 \text{var}(\eta)}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 \text{var}(\eta) + \sum_{j=1}^{J} \text{var}(\varepsilon_j)}, \quad (2)$$

where $\rho_s$ is the reliability of the scores of a subscale, $\lambda_j$ are the factor loadings of the items of the subscale, var($\eta$) is the variance of the subscale's factor, and var($\varepsilon_j$) are the error variances of the subscale's items. The formula presented above is adequate to estimate the reliability of scores of scales composed of testlets because the error variances estimated with the CTCU model contain nonsystematic effects as well as unmeasured method effects (Lance et al., 2002). This formula does not include the covariances between errors because the CTCU model fit to the data only allowed errors of indicators of different factors to covary (see Figure 1). The reliability estimates for the scores of the VARK subscales were .85, .82, .84, and .77 for the visual, aural, read/write, and kinesthetic subscales, respectively, which are considered adequate.

## Discussion

This study aimed to produce some evidence of whether the scores of the VARK learning style inventory support the four-factor structure of the scale hypothesized by its author. An analysis of the dimensionality of the VARK is important to help validate

the VARK scores' use as a learning styles diagnostic tool. As the very large number of observations collected in just 1 month indicates, the VARK has been a very popular tool for identifying learning styles. We found that a four-factor CTCU model fits the observed data and that the reliability estimates of the scores of the VARK were adequate. However, these conclusions can be considered preliminary because not all fit indices supported the fit of the model and the estimated factor loadings of the VARK items were small to moderate. These limitations may be because of a variety of factors, such as latent constructs that are not accounted for by the model and/or unobserved mixtures of populations. Additional studies are needed to strengthen the evidence of model fit, and revision or removal of certain items may be necessary to ensure the validity of the VARK scores. Possible ways to improve model fit include specification searches (Marcoulides & Drezner, 2001, 2003) to fine tune the measurement model underlying VARK's scores and item selection methods (Leite, Huang, & Marcoulides, 2008) to identify which VARK items maximize model fit, reliability, and relationships with external criteria. Furthermore, factor mixture models (Bauer & Curran, 2004; Lubke & Muthén, 2005) could be used to evaluate whether the VARK scores come from a mixture of populations with different factor structures of learning styles.

Lee et al. (2000) discuss the difference between fixed and random testlets, which is related to the difference between a scale and a form of a scale. For example, the MCAT is a scale with different forms that are constantly being developed. If the construction of each form of scale is based on randomly sampling content to form testlets, the testlets can be considered random. Lee, Dunbar, and Frisbie (2001) argued that with random testlets, the CTCU model would be more appropriate than the CTCM model. Because the VARK currently has only one form and its testlets were purposefully chosen, the testlets were conceptualized as fixed in this study. If in the future different forms of the VARK are developed based on sampling of learning scenarios, an analysis of method effects using multilevel MTMM-CFA would be possible.

The VARK has evolved over its lifetime as items are refined and rewritten to improve the scale scores' validity and reliability. However, we would like to recognize that the language of the items is occasionally difficult to interpret consistently across a wide range of respondents. For example, items sometimes seem to be more a reflection of the preferences of the audience the respondent is hypothetically addressing than the respondent himself or herself. Consider the item

> You are planning a holiday for a group. You want some feedback from them about the plan. You would: 1. Phone, text or email them; 2. Describe some of the highlights; 3. Give them a copy of the printed itinerary; 4 use a map or website to show them the places.

In this item, is the choice a reflection of the respondent's preference or his or her interpretation of what the audience needs? In addition, items seem to be directed at a specific audience with a rather high level of economic options, thus limiting its usability as a universal instrument. For example, items sometimes refer to purchasing

a digital camera or mobile phone, making a speech at a conference or special occasion, having a preference for Web sites with different characteristics, and so on. The users who could reliably respond to these items represent only a subsection of the population at large. It would be useful to note this restriction in any situation where the VARK is being used; is the population for which it is being used similar enough to the population for which it was developed to be able to define the same subsets of preferences as intended? To what extent and in what ways do these characteristics affect the reliability and validity of the resulting scores?

The preliminary evidence of validity of the VARK scores with respect to dimensionality and reliability found in the current study support the use of the VARK as a low-stakes diagnostic tool by students and teachers. Therefore, those who wish to use the instrument as a way of helping students identify their preferences should feel comfortable in this use. The large amount of material provided on the VARK Web site to help learners adapt their learning strategies to materials representing different modes of presentation are definitely useful, and students should be encouraged to explore them.

A second use of the VARK that has been considered is as a research instrument whose scores could serve as predictors or outcomes for the evaluation of instructional methods. Although the information about dimensionality and reliability of the VARK scores reported in this paper are important pieces of evidence of validity, they are not sufficient to support the use of the VARK with research. Before the VARK is used with research, other sources of evidence of the validity of its scores should be collected. In this study, we focused on validity evidence with respect to VARK's internal structure. However, validity evidence with respect to content, response process, relations with other variables (i.e., convergent, discriminant, and test-criterion relationships), and consequences of testing should be obtained to build a strong argument for the validity of VARK scores' use with research. For example, convergent evidence of validity could emerge from examining the relationship between examinees' scores on VARK factors and their scores on factors measured by other learning styles scales. Evidence with respect to test-criterion relationships could be obtained by contrasting examinees' learning style preferences measured by the VARK with their performance on tasks specifically developed to use predominantly a single learning style. Furthermore, validity evidence with respect to response process could be obtained by performing cognitive interviews (Willis, 1999) with examinees during or after the completion of the VARK.

Given that the VARK aims to measure four latent traits, validity arguments must be provided for interpretations of these traits (Kane, 2006). However, no validity information is available for VARK's learning styles classification algorithm implemented through the VARK Web site. The VARK algorithm classifies examinees into having very strong, strong, or mild single learning style preferences or any combination of two, three, or four learning style preferences. Because the examinees taking the VARK are allowed to select more than one option within a question, the prevalence of examinees classified as having multiple learning styles is high. For example, with our sample, 35.8% of the examinees was classified by the algorithm as having all four learning style preferences, while the percentage of examinees with a single learning

style preference was only 29.7%. Therefore, the separation of individuals between a group with multiple preferences and a group with a single preference is more pronounced than the separation of individuals with different learning styles. Kane (2006) suggested that evidence for the scoring inference of scales can be obtained based on expert panels. Therefore, the learning style classifications provided by the VARK algorithm could be examined against the judgment of experts in learning styles. The meaning of very strong, strong, or mild learning style preferences could be clarified by reference to performance on tasks that demand a single or many learning styles. Furthermore, multiple-group CFA could be used to verify whether the groups defined by VARK's scoring algorithm present substantial mean differences in latent traits.

One limitation of the current analysis is that the data were obtained only through Internet test administration. Although the sample size was very large, it is possible that the use of the Internet to collect data resulted in an exaggerated number of responses from individuals who are very skilled with the computer. Because the Internet presents stimuli in all four perceptual modes, it is possible that the proportion of respondents with preference for all learning styles in this sample is thus inflated. Although this paper provided some evidence of the validity of the VARK scores, it also indicates that researchers using the VARK should proceed with caution because the use and proposed interpretations of VARK scores have not yet received a comprehensive validation.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

## Funding

The authors received no financial support for the research and/or authorship of this article.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson, J. A., & Adams, M. (1992). Acknowledging the learning styles of diverse student populations: Implications for instructional design. In L. L. B. Border & N. V. N. Chism (Eds.), *Teaching for diversity* (pp. 19-34). San Francisco: Jossey-Bass.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9,* 3-29.

Cassidy, S. (2004). Learning styles: An overview of theories, models, and measures. *Educational Psychology, 24,* 419-444.

Dunn, R., Beaudry, J., & Klavas, A. (1989). Survey of research on learning styles. *Educational Leadership, 46,* 50-59.

Dunn, R., Dunn, K., & Price, G. (1985). *Manual: Learning style inventory*. Lawrence, KS: Price Systems.

Dunn, R., Griggs, S. A., Olson, J., & Beasly, M. (1995). A meta-analytic validation of the Dunn and Dunn model of learning style preferences. *Journal of Educational Research, 88,* 353-362.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65,* 241-261.

Eid, M., Lischetzke, T., Nussbeck, F., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8,* 38-60.

Fleming, N. D. (2001). *Teaching and learning styles: VARK strategies*. Christchurch, New Zealand: Author.

Fleming, N. D. (2008). *The VARK Questionnaire*. Retrieved December 29, 2008, from http://www.vark-learn.com/english/page.asp?p=questionnaire

Fleming, N. D., & Mills, C. (1992). Not another inventory, rather a catalyst for reflection. *To Improve the Academy, 11,* 137-143.

Henson, R. K., & Hwang, D. (2002). Variability and prediction of measurement error in Kolb's Learning Style Inventory scores: A reliability generalization study. *Educational and Psychological Measurement, 62,* 712-727.

Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and instruction*. Hillsdale, NJ: Lawrence Erlbaum.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136-153.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Keefe, J. W. (1987). *Learning style: Theory and practice*. Reston, VA: National Association of Secondary School Principals.

Kolb, D. A. (1976). Management and the learning process. *California Management Review, 18,* 21-31.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on coefficient alpha. *Applied Psychological Measurement, 21,* 337-348.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method (CTCM) and correlated uniqueness (CU) models for multitrait–multimethod (MTMM) data. *Psychological Methods, 7,* 228-244.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 9-15.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement, 61,* 958-975.

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research, 43,* 411-431.

Lubke, G., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21-39.

Marcoulides, G. A., & Drezner, Z. (2001). Specification searches in structural equation modeling with a genetic algorithm. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 247-268). Mahwah, NJ: Lawrence Erlbaum.

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling, 10,* 154-164.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335-361.

Marsh, H. W., Hau, K., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275-340). Mahwah, NJ: Lawrence Erlbaum.

Muthén, L & Muthén, B. O. (2006). *Mplus (Version 4.2).* Los Angeles: Author.

Muthén, L., & Muthén, B. O. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles: Author.

Raykov, T. (2001a). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25,* 69-76.

Raykov, T. (2001b). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54,* 315-323.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Riding, R. (2001). Nature and effects of cognitive style. In R. Sternberg & L.-F. Zhang (Eds.), *Perspectives on thinking, learning, and cognitive styles* (pp. 47-72). Mahwah, NJ: Lawrence Erlbaum.

Sternberg, R., & Zhang, L.-F. (Eds.). (2001). *Perspectives on thinking, learning, and cognitive styles*. Mahwah, NJ: Lawrence Erlbaum.

Thelen, H. A. (1954). *Dynamics of groups at work.* Chicago: University of Chicago Press.

Tomás, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35,* 469-499.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203-220.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29,* 126-149.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1-26.

Willis, G. B. (1999). *Cognitive interviewing: A "how to" guide*. Research Triangle Park, NC: Research Triangle Institute.