Measuring Self-Efficacy: Multitrait-Multimethod Comparison of Scaling Procedures

Mimi Bong a; Dennis Hocevar b

a Department of Educational Psychology, University of South Carolina. b Division of Learning and Instruction, University of Southern California.

Online Publication Date: 01 April 2002

To cite this Article Bong, Mimi and Hocevar, Dennis(2002)'Measuring Self-Efficacy: Multitrait-Multimethod Comparison of Scaling Procedures'.Applied Measurement in Education,15:2,143 — 171

To link to this Article: DOI: 10.1207/S15324818AME1502_02

URL: http://dx.doi.org/10.1207/S15324818AME1502_02

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Measuring Self-Efficacy:
Multitrait-Multimethod Comparison of Scaling Procedures

Mimi Bong
Department of Educational Psychology
University of South Carolina

Dennis Hocevar
Division of Learning and Instruction
University of Southern California

Convergent and discriminant validity of various self-efficacy measures was examined across 2 studies. In Study 1, U.S. high school students \((N = 358)\) rated their self-efficacy in 6 school subjects in reference to either specific problems or general self-efficacy statements on the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1990). In Study 2, Korean female high school students \((N = 235)\) judged their perceived efficacy in reference to specific problems, specific task descriptions, and MSLQ statements in 3 school subjects. Across Studies 1 and 2, the 1st-order confirmatory factor analyses (CFAs) provided support for both convergent validity of different self-efficacy responses and discriminant validity of perceived self-efficacy across different subject areas. The 2nd-order CFAs confirmed the discriminant validity of self-efficacy beliefs. Substantial method effects were also observed. The problem- and task-referencing methods correlated with each other to a greater extent than they did with the MSLQ Self-Efficacy scale.

The primary purpose of this investigation was to assess the equivalence of self-efficacy judgments that were measured by different methods. Convergent and discriminant validity of academic self-efficacy responses were examined in a multitrait-multimethod (MTMM) framework. In Study 1, U.S. high school students reported their self-efficacy perceptions in six school subjects by rating either their
confidence for solving specific problems presented or their agreement with each of
the self-efficacy statements provided. In Study 2, Korean high school students re-
ported their self-efficacy in three school subjects in reference to either specific
problems, written task descriptions, or general self-efficacy statements. Confirm-
atory factor analyses (CFAs) and higher order confirmatory factor analyses
(HCFAs) were applied to these MTMM self-efficacy data.

BRIEF OVERVIEW OF SELF-EFFICACY RESEARCH

Self-efficacy refers to one’s convictions to successfully organize and execute a
course of action that is required to achieve a desirable outcome (Bandura, 1997). It
is context-specific judgment that is closely tied to the specific domain and situation
in question (Zimmerman, 1995). Academic self-efficacy, in particular, represents
learners’ subjective confidence for successfully performing given academic tasks
at designated levels (Schunk, 1991). As such, it wields a critical influence on virtu-
ally all aspects of student learning. Students with a strong sense of self-efficacy
willingly choose challenging academic tasks (Bandura & Schunk, 1981), use ef-
effective learning strategies (Pintrich & De Groot, 1990), persist longer in the face of
difficulties (Lent, Brown, & Larkin, 1984), and set higher academic goals (Zim-
merman, Bandura, & Martinez-Pons, 1992). These students also demonstrate
more positive attitudes and emotions toward learning as evidenced by their higher
academic aspirations and lower depression (Bandura, Barbaranelli, Caprara, &
Pastorelli, 1996), lower anxiety (Pajares & Miller, 1994), and lower apprehension
(Pajares, Miller, & Johnson, 1999) in academic contexts. Through its positive in-
fluence on subsequent motivation and learning, heightened self-efficacy brings
about better academic performance and self-regulation (Multon, Brown, & Lent,

As evidence demonstrating the potency of academic self-efficacy beliefs accu-
mulates, an increasing number of researchers have incorporated this important
construct in their investigations. Up until now, operational definitions of self-effi-
cacy have been relatively more consistent compared to those of other self-con-
structs (Bong & Clark, 1999). Still, they are not without some discrepancy. Investi-
gators have used different methods to assess self-efficacy perceptions, which
sometimes renders comparability of the findings unclear. In part, the problem reso-
nates Pajares’ (1996) comment on the specificity of self-efficacy beliefs and their
correspondence to criterial tasks. The outcome of interest in educational research
ranges from performance on a very specific task to more general-level indicators
such as course choice or semester grades. Because outcomes like course grades
typically reflect some form of aggregation of students’ performances on diverse
tasks and activities, they pose additional complexity to self-efficacy assessment.
Given the growing trend in self-efficacy research, it is important to evaluate different assessment methods in relation to the self-efficacy theory.

MEASURING SELF-EFFICACY

There have been four broad categories of measurement techniques that researchers use to assess the strength of self-efficacy beliefs. The first and standard method of measuring academic self-efficacy is to present a set of specific problems, performance on which is the very target of prediction. Students report their confidence for successfully solving each type of problems on a scale ranging from 0 to 100 with a 10-unit interval. The following verbal descriptors usually accompany the scale: 10 (not sure), 40 (somewhat sure), 70 (pretty sure), and 100 (very sure). Schunk and his colleagues (Schunk, 1982, 1983; Schunk & Cox, 1986; Schunk & Gunn, 1986; Schunk & Hanson, 1985, 1989; Schunk, Hanson, & Cox, 1987) repeatedly used this method for measuring elementary school students’ arithmetic self-efficacy. Zimmerman and Martinez-Pons (1990), in their comparison of gifted and regular school students’ self-efficacy perceptions, presented verbal (i.e., word defining) and math problems of increasing difficulty (i.e., simple arithmetic, algebra, probability, and statistics) and asked students to rate their perceived capability to solve each of the problems. Zimmerman and Kitsantas (1999) likewise obtained students’ self-efficacy ratings for a sentence-combining task by presenting specific writing revision problems.

The second category of self-efficacy assessment method is similar to the first category in that it provides concrete anchors that respondents use for gauging their efficacy perceptions. However, what are being presented are not specific problems but verbal descriptions of specific task components that reflect the major aspects of successful performance. Researchers choose this method when the target performance cannot easily be summed up as specific problems. In reading, for example, students are asked to judge their confidence to successfully perform tasks such as read one of the textbooks; know all the words on a page in one of the schoolbooks; know the meaning of plurals, prefixes, and suffixes; and understand the main idea of a story (Shell, Colvin, & Bruning, 1995). In writing, students estimate their confidence for performing such tasks as write a one-page summary of a book; correctly punctuate a sentence; correctly spell all words in a one-page story or composition; and correctly use parts of speech such as nouns, verbs, adjectives, or adverbs (Pajares et al., 1999; Shell et al., 1995). This method is also used often to describe computer-related skills such as using Hypercard clip art; creating a background design that is used by multiple cards (Schunk & Ertmer, 1999), downloading necessary materials from the Web; and using Internet search engines such as Yahoo (Joo, Bong, & Choi, 2000).

Whereas these first two methods concentrate on specific facets of task performance, the latter two methods concentrate more on the overall performance levels.
One of them is to ask students about their confidence to achieve a specific letter grade. Students rate the strength of their beliefs that they could obtain each of the letter grades ranging from A to F (Zimmerman & Bandura, 1994). The other method is to ask students to judge their general confidence to function successfully in the given domain without making an explicit reference to any individual problems or tasks. Instead, descriptions of generic tasks that are commonly performed in most academic domains are provided in the context of specific subjects. Therefore, students rate how much they agree with statements like “I am certain that I can understand what is taught in (a specific subject) class,” “I expect to do very well in (a specific subject) class,” and “I am certain that I can figure out how to do the most difficult schoolwork in (a specific subject)” (Pintrich & De Groot, 1990).

To date, several researchers have addressed the issue of self-efficacy scale differences in terms of predictive validity. For example, Pajares and Miller (1995) presented convincing evidence that how one assesses self-efficacy judgment could produce different results regarding its relations with relevant outcomes. The researchers solicited college students’ confidence ratings for solving specific math problems, completing everyday math tasks, or performing successfully in math-related courses. As expected, math problems self-efficacy was a better predictor of math problem-solving performance than math courses self-efficacy. Math courses self-efficacy predicted choice of math-related majors better than math problems self-efficacy. The three self-efficacy scores were all highly correlated among themselves as well as with the two outcome measures. Bong (2001b) also compared multiple self-efficacy scores that were assessed at varying levels of specificity. College students reported their confidence for correctly solving specific problems presented, successfully mastering the representative topics of the course, successfully performing in the course, and performing well in college courses in general. As Pajares and Miller (1995) observed, all self-efficacy scores were positively correlated among themselves and, with an exception of problem-specific self-efficacy, with the value students perceived in the course. More interesting, correlation between any two self-efficacy scores decreased as the difference in their measurement levels increased.

These investigations are instrumental in establishing the basic guidelines for assessing self-efficacy beliefs. The positive correlation among different self-efficacy scales reported in both studies also provides some evidence of convergent validity. However, issues of convergent and discriminant validity of scales can be dealt with more effectively in a MTMM design (Campbell & Fiske, 1959), which requires a minimum of two traits assessed by at least two methods. Because most previous

---

1The term trait is a misnomer for the study of self-efficacy. Bandura (1997), as well as many self-efficacy researchers, made it clear that self-efficacy is a context-specific judgment that should not be viewed as one of the personality traits. We decided to retain this term simply to avoid any conceptual difficulty that may arise from using a different term for the well-established MTMM procedures.
studies measured self-efficacy perceptions in relation to a single domain, convergent and discriminant validity of self-efficacy responses have not been probed systematically according to the Campbell and Fiske criteria. In this research, multiple measures of self-efficacy beliefs in multiple academic domains were available across two studies, allowing MTMM comparison of self-efficacy scores. Unlike traditional MTMM analysis, this study applied CFA and HCFA. CFA is especially useful in situations in which linkages between observed variables and latent constructs can be clearly established according to the theory. Because most measures in social and behavioral sciences contain sizable measurement errors, CFA affords important advantages over a zero-order correlation approach. Rather than relying on an unrealistic assumption that the measures are perfect, CFA takes the measurement errors into account. In the MTMM context, it also allows partitioning of the indicator variance into trait, method, and random error components. Researchers generally agree that CFA is the most defensible and informative approach to the analysis of the MTMM (Marsh, 1993; Marsh & Hocevar, 1983). By applying CFA and HCFA procedures to MTMM self-efficacy matrices, this study aimed at examining (a) the equivalence of self-efficacy responses from different assessment methods (i.e., convergent validity) and (b) the distinctiveness of self-efficacy beliefs in different academic domains (i.e., discriminant validity).

STUDY 1

Method

Participants

The sample consisted of 358 students (49% boys) enrolled in four high schools in Los Angeles county at the time of the survey. Among the 588 students who participated in the larger research project (see Bong, 1997b, for a description of the larger sample), students who reported having previous experience with all six subject areas were selected. Ethnic composition of our sample was 16% White, 6% African American, 55% Hispanic, 21% Asian, and 2% Native American and other. Students were mostly in Grades 11 (21%) and 12 (78%).

Measures and Procedures

Problem-referenced self-efficacy. Seven typical problems from six school subjects (i.e., English, Spanish, American history, algebra, geometry, and chemistry) were prepared from the Scholastic Aptitude Test I and II preparatory booklets (Brownstein, Weiner, & Green, 1994; College Entrance Examination Board and Educational Testing Service, 1994; see Bong, 1997b, for sample problems). Care was taken to ensure that problems of representative types and moderate difficulty were included. Each problem was presented through an overhead projector for a
duration that was long enough to recognize its type but too short to attempt its solution. Students rated how confident they were to correctly solve the types of problems presented on a scale ranging from 0 to 100 in 10-unit intervals. The following verbal descriptors were provided to help students understand more clearly what each number represented: 0 (not sure), 40 (maybe), 70 (pretty sure), and 100 (real sure). This is a standard procedure of assessing self-efficacy beliefs using specific problems (see, e.g., Bandura, 1997, pp. 42–46). One might argue that the difference in response scales (e.g., using a 0–100 scale vs. using a 1–7 scale; see following) could confound the results regarding the different types of measurement. Although this certainly is a possibility, we felt that using the most typical assessment strategies associated with each method would provide more insights as to the difference in measurement methods as they are being used in the current literature.

Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1990) self-efficacy. Students responded to self-efficacy items on the MSLQ. The MSLQ self-efficacy items seek students’ endorsement ratings on statements describing general academic events in the context of specific domains. Of the nine items on the original scale, three ask students to compare their capability to that of their peers. Self-efficacy researchers maintain that judgments of self-efficacy depend more heavily on the mastery criteria (i.e., being able to succeed) than on the normative ones (i.e., being better than others; Bong & Clark, 1999; Zimmerman, 1995, 1996). Accordingly, comparative items were excluded from this investigation. The final scale contained the following six items for each school subject: “I’m certain that I can understand what is taught in (a specific school subject) class,” “I expect to do very well in (a subject) class,” “I am sure that I can do an excellent job on the problems and tasks assigned for (a subject) class,” “I know that I will be able to learn the material for (a subject) class,” “My study skills are excellent in (a subject) class,” and “I think I will receive a good grade in (subject) class.” Response categories ranged from 1 (not at all true) to 7 (very true), as in Pintrich and De Groot (1990).

Results

Table 1 presents descriptive statistics of the scales. There were 42 items in the problem-specific self-efficacy scale (i.e., 7 items × 6 subjects) and 36 items in the MSLQ scale (i.e., 6 items × 6 subjects). One of the early decisions that need to be made in CFA is whether to perform the analyses with individual items or item parcels, which can be created by combining responses to multiple items. In both Studies 1 and 2, we decided to use item parcels rather than individual items as indicators for several reasons. First, the sample sizes did not permit such elaborate analyses of using single-item indicators. We acknowledge that the probability of obtaining proper solutions improves with a larger number of indicators per factor when sample size is
small and when individual items are used as indicators (Marsh, Hau, Balla, & Grayson, 1998). When using three or more item parcels, however, solutions almost always converge and are unaffected by the sample size. Second, item parcels are known to meet the multivariate normality assumption that underlies structural equation modeling better than individual items (Kline, 1998, p. 237). Third and perhaps most important, the scales used in this research appeared reasonably unidimensional so as not to distort the results with the use of item parceling.

Bandalos and Finney (2001) suggested that the use of item parcels may lead to misspecification of models and biases in parameter estimates. They stated that

The crucial factors in a researcher’s decision to use item parceling appears to be the degree to which he or she is willing to make the assumption that the use of item par-

<table>
<thead>
<tr>
<th>Study and Scale</th>
<th>No. Items</th>
<th>Item M</th>
<th>Item SD</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1 Problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>7</td>
<td>74.713</td>
<td>21.982</td>
<td>.850</td>
</tr>
<tr>
<td>Spanish</td>
<td>7</td>
<td>71.173</td>
<td>32.250</td>
<td>.963</td>
</tr>
<tr>
<td>History</td>
<td>7</td>
<td>67.981</td>
<td>24.348</td>
<td>.905</td>
</tr>
<tr>
<td>Algebra</td>
<td>7</td>
<td>66.441</td>
<td>27.535</td>
<td>.915</td>
</tr>
<tr>
<td>Geometry</td>
<td>7</td>
<td>64.505</td>
<td>28.162</td>
<td>.924</td>
</tr>
<tr>
<td>Chemistry</td>
<td>7</td>
<td>55.279</td>
<td>28.656</td>
<td>.891</td>
</tr>
<tr>
<td>MSLQ English</td>
<td>6</td>
<td>5.512</td>
<td>1.369</td>
<td>.891</td>
</tr>
<tr>
<td>Spanish</td>
<td>6</td>
<td>4.850</td>
<td>1.804</td>
<td>.951</td>
</tr>
<tr>
<td>History</td>
<td>6</td>
<td>5.393</td>
<td>1.366</td>
<td>.926</td>
</tr>
<tr>
<td>Algebra</td>
<td>6</td>
<td>4.829</td>
<td>1.627</td>
<td>.953</td>
</tr>
<tr>
<td>Geometry</td>
<td>6</td>
<td>4.473</td>
<td>1.663</td>
<td>.962</td>
</tr>
<tr>
<td>Chemistry</td>
<td>6</td>
<td>4.174</td>
<td>1.643</td>
<td>.966</td>
</tr>
<tr>
<td>Study 2 Problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>15</td>
<td>69.688</td>
<td>16.583</td>
<td>.957</td>
</tr>
<tr>
<td>English</td>
<td>25</td>
<td>65.550</td>
<td>17.005</td>
<td>.974</td>
</tr>
<tr>
<td>Math</td>
<td>25</td>
<td>52.999</td>
<td>21.421</td>
<td>.981</td>
</tr>
<tr>
<td>Tasks Korean</td>
<td>10</td>
<td>61.944</td>
<td>16.207</td>
<td>.937</td>
</tr>
<tr>
<td>English</td>
<td>10</td>
<td>59.293</td>
<td>17.269</td>
<td>.962</td>
</tr>
<tr>
<td>Math</td>
<td>10</td>
<td>61.068</td>
<td>17.709</td>
<td>.912</td>
</tr>
<tr>
<td>MSLQ Korean</td>
<td>5</td>
<td>3.154</td>
<td>.760</td>
<td>.872</td>
</tr>
<tr>
<td>English</td>
<td>5</td>
<td>3.096</td>
<td>.804</td>
<td>.910</td>
</tr>
<tr>
<td>Math</td>
<td>5</td>
<td>3.152</td>
<td>.802</td>
<td>.910</td>
</tr>
</tbody>
</table>

Note. MSLQ = Motivated Strategies for Learning Questionnaire.
ceiling has not masked any substantively and/or theoretically important sources of lack of fit. These sources of lack of fit typically, in CFA applications, involve some sort of unmodeled minor or secondary factor, which results in covariances among the uniquenesses or may result from items that load on more than one factor. (Bandalos & Finney, 2001, p. 286)

They thus urged researchers to model the data pattern at the item level first so that possible sources of misspecification could be located. Researchers can then determine whether the unmodeled sources of variance are ignorable or should be explicitly incorporated in the model.

Following Bandalos and Finney’s (2001) recommendation, we report results of the item-level analyses before reporting on the full MTMM models that used item parcels. The purpose of these item-level analyses was to ensure the unidimensionality of scales, which is a requisite condition for creating item parcels. A CFA model at the item level with six problem-specific self-efficacy factors (i.e., English, Spanish, History, Algebra, Geometry, and Chemistry Problem Self-Efficacy) with no cross loading and no correlated errors did not fit the data well, $\chi^2$ (804, $N = 358$) = 2,827.742, $p < .001$ (nonnormed fit index [NNFI] = .807, comparative fit index [CFI] = .820). Results of the Lagrange Multiplier tests (Bentler, 1995) suggest that correlation paths be opened between various error terms, mostly within the boundary of each scale. Incorporating 33 error covariances (i.e., 2 in English, 15 in history, 10 in Spanish, and 4 in geometry) improved model fit to the marginally acceptable level, $\chi^2$ (771, $N = 358$) = 1,833.089, $p < .001$ (NNFI = .895, CFI = .906). Eight of these error correlations involved items that were consecutively presented within each academic domain. One correlation was between items that dealt with aspects of English grammar. Overall, there was no particularly strong or consistent pattern among these error covariances and most of them appeared random. We thus concluded that these effects were best regarded as such. With the MSLQ data, an item-level CFA model with six self-efficacy factors demonstrated acceptable fit, $\chi^2$ (579, $N = 358$) = 1,581.462, $p < .001$ (NNFI = .926, CFI = .932). Modification indexes nonetheless suggested that specifying correlated error paths between items sharing the same wording would significantly improve the model fit. A revised model with these additional parameters fit the data better, $\chi^2$ (489, $N = 358$) = 995.037, $p < .001$ (NNFI = .956, CFI = .966). Although these covariances did not necessarily challenge the unidimensionality of MSLQ scales, we felt that they were nonetheless systematic variance that needed to be accounted for. These effects were incorporated indirectly in all subsequent models in the form of correlated errors between item parcels containing the identically worded items.

As a way of examining the possibility of parameter biases due to item parceling, we compared the factor correlation coefficients that were obtained from the final within-scale, item-level models with those from the full MTMM model that used item parcels. For the problem-specific self-efficacy scales, magnitude of these dif-
ferences ranged from 0 to .188. The average difference was .057. The largest difference was between the English and History Problem Self-Efficacy factors (.424 in the item-level model and .612 in the parcel-level model). For the MSLQ scales, the differences ranged from .001 to .068. The average difference was .022. Other than the few differences in the problem-specific scales that were greater than .100, factor correlations appeared remarkably similar. After carefully examining potential sources of misspecification at the item level, we decided that these effects were mostly random and were of no substantive or theoretical importance. Using item parcels also reduced the computer resources necessary for model run. To minimize the source of confounding when comparing the scales, we used item parcels for all scales.

Eighteen measured variables (MVs) were created for Problem Self-Efficacy by combining responses to two to three problems in each academic domain. Specifically, responses to Problems 1, 4, and 7, Problems 2 and 5, and Problems 3 and 6 in each subject were averaged to produce three MVs for each of the six school subjects. Another 18 MVs were created for MSLQ Self-Efficacy by combining responses to Items 1 and 4, Items 2 and 5, and Items 3 and 6 in each domain (descriptive statistics of individual items and MVs for both scales are available from Mimi Bong). Therefore, there were 6 MVs in each school subject, 3 of which shared the same method. The problem referencing and the MSLQ were treated as two methods, whereas self-efficacy perceptions in the six school subjects were treated as six traits. The pattern of covariation among these MVs was likely created by many factors, most notably by the method and trait effects. Depending on the relative contribution of each source, the number of factors required to obtain satisfactory model fit would differ. Different CFA models were thus specified and compared. All CFAs were conducted with the EQS program (Bentler, 1995). Because the two methods used very different response scales, we conditioned the Problem Self-Efficacy matrix by dividing all responses by 10.

If observed variation among MVs was mostly due to method effects (i.e., students provided similar ratings to self-efficacy items using the same method regardless of the content domain being tapped), two correlated method factors alone should be able to illustrate the data to a sufficient degree (Model 1). On the other hand, if the data pattern was created mostly because of the different traits (i.e., students’ self-efficacy ratings were primarily determined by their self-efficacy beliefs in the subject domain, irrespective of the assessment tools), six correlated trait factors should suffice (Model 2). The NNFI and CFI reported in Table 2 represent improvement in fit of the hypothesized model in comparison with a null model and, as such, are indicators of the overall model fit. Values greater than .90 are commonly taken as evidence of satisfactory model fit. Neither Model 1 (NNFI = .413, CFI = .490) nor Model 2 (NNFI = .709, CFI = .753) was able to reproduce the observed data to a satisfactory degree.

Model 3 specified six correlated traits and two correlated method factors. All fit indexes improved substantially, falling only little short of the recommended cutoff
value of .90 (NNFI = .872, CFI = .890; see Table 2). This strongly attests to the need for both trait and method factors. Unfortunately, there is inherent danger of partial underidentification when one is dealing with only two methods (Marsh & Hocevar, 1983). Model 4 avoided this problem by specifying 12 first-order factors, each of which represents a unique combination of a particular trait and method (e.g., English self-efficacy assessed by the problem-referencing method). It also allowed examining the validity issue according to the Campbell and Fiske (1959) guidelines. Model fit improved substantially with fit indexes well above the acceptable value (NNFI = .970, CFI = .977). Each factor was clearly defined with sizable standardized factor loadings ranging from .776 to .960 (\(Mdn = .902\)). Table 3 presents correlation coefficients among the 12 factors.

Factor correlation coefficients among the six Problems Self-Efficacy factors ranged from .085 to .881, whereas those among the MSLQ Self-Efficacy factors ranged between .124 and .830. Within each method, the highest correlation emerged between Algebra and Geometry Self-Efficacy factors. Because these two factors were very highly correlated, three additional CFA models were run to test whether they could be combined into a single Math Self-Efficacy factor. Model 5 specified 11 factors by combining Problem Algebra and Problem Geometry factors of Model 4 into a single Problem Math Self-Efficacy factor. Model 6 likewise hypothesized 11 factors by combining MSLQ Algebra and MSLQ Geometry factors into an MSLQ Math factor. Model 7 combined Algebra and Geometry

### Table 2

Goodness-of-Fit Indexes of Confirmatory Factor Analysis
Models Tested in Study 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>(\chi^2)</th>
<th>df</th>
<th>NNFI</th>
<th>CFI</th>
<th>Res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>36 uncorrelated first-order factors</td>
<td>15062.164</td>
<td>630</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>2 correlated method factors only</td>
<td>7911.842</td>
<td>548</td>
<td>.413</td>
<td>.490</td>
<td>.089</td>
</tr>
<tr>
<td>2</td>
<td>6 correlated trait factors only</td>
<td>4096.694</td>
<td>534</td>
<td>.709</td>
<td>.753</td>
<td>.099</td>
</tr>
<tr>
<td>3</td>
<td>6 correlated trait and 2 correlated method factors with no trait-method correlation</td>
<td>1857.165</td>
<td>497</td>
<td>.881</td>
<td>.906</td>
<td>.060</td>
</tr>
<tr>
<td>4</td>
<td>12 correlated trait–method combination factors</td>
<td>816.562</td>
<td>483</td>
<td>.970</td>
<td>.977</td>
<td>.023</td>
</tr>
<tr>
<td>5</td>
<td>Problems Algebra and Problems Geometry factors in Model 4 are combined into Problems Math</td>
<td>975.679</td>
<td>494</td>
<td>.957</td>
<td>.967</td>
<td>.024</td>
</tr>
<tr>
<td>6</td>
<td>MSLQ Algebra and MSLQ Geometry factors are combined into MSLQ Math</td>
<td>1323.203</td>
<td>494</td>
<td>.927</td>
<td>.943</td>
<td>.025</td>
</tr>
<tr>
<td>7</td>
<td>Both Problems and MSLQ Algebra and Geometry factors are combined into Problems Math and MSLQ Math</td>
<td>1456.232</td>
<td>505</td>
<td>.918</td>
<td>.934</td>
<td>.027</td>
</tr>
</tbody>
</table>

**Note.** NNFI = Bentler–Bonnett nonnormed fit index; CFI = comparative fit index; Res. = average absolute standardized residuals; MSLQ = Motivated Strategies for Learning Questionnaire.
### TABLE 3

Factor Correlations of Model 4 in Study 1

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Problem SE</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish Problem SE</td>
<td>.085</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History Problem SE</td>
<td>.612</td>
<td>.243</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algebra Problem SE</td>
<td>.520</td>
<td>.106</td>
<td>.419</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometry Problem SE</td>
<td>.482</td>
<td>.114</td>
<td>.463</td>
<td>.881</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemistry Problem SE</td>
<td>.535</td>
<td>.195</td>
<td>.699</td>
<td>.714</td>
<td>.671</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English MSLQ SE</td>
<td>.501</td>
<td>-.072</td>
<td>.319</td>
<td>.189</td>
<td>.107</td>
<td>.220</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish MSLQ SE</td>
<td>.003</td>
<td>.723</td>
<td>.117</td>
<td>.084</td>
<td>.074</td>
<td>.138</td>
<td>.169</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History MSLQ SE</td>
<td>.368</td>
<td>-.008</td>
<td>.474</td>
<td>.214</td>
<td>.189</td>
<td>.291</td>
<td>.601</td>
<td>.124</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algebra MSLQ SE</td>
<td>.284</td>
<td>.122</td>
<td>.219</td>
<td>.646</td>
<td>.584</td>
<td>.424</td>
<td>.251</td>
<td>.311</td>
<td>.353</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometry MSLQ SE</td>
<td>.287</td>
<td>.135</td>
<td>.282</td>
<td>.613</td>
<td>.668</td>
<td>.507</td>
<td>.260</td>
<td>.294</td>
<td>.342</td>
<td>.830</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Chemistry MSLQ SE</td>
<td>.352</td>
<td>.161</td>
<td>.391</td>
<td>.520</td>
<td>.497</td>
<td>.659</td>
<td>.274</td>
<td>.306</td>
<td>.386</td>
<td>.644</td>
<td>.741</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Note.* Bold entries represent monotrait heteromethod coefficients; italic entries represent heterotrait monomethod coefficients. SE = Self-Efficacy; MSLQ = Motivated Strategies for Learning Questionnaire.
Self-Efficacy factors of both Problem and MSLQ methods into Problem Math and MSLQ Math factors, thus specifying only 10 trait–method combination factors. As can be seen in Table 2, in all three instances, combining algebra and geometry indicators to load on the same Math factor resulted in poorer overall model fit compared to Model 4. Loadings of relevant indicators also showed a uniform decline when only a single Math Self-Efficacy factor was specified instead of separate Algebra and Geometry Self-Efficacy factors. Correlation coefficients among other self-efficacy factors within each method were substantially less than 1.0. These results demonstrate discriminant validity of the six subject-specific self-efficacy factors as assessed by each method.

In the first-order CFA, evidence of convergent validity can be found when (a) statistically significant and substantial loadings on the trait factors are obtained and (b) significant decrement in fit is observed when trait factors are deleted from model specification (Gardner, Cummings, Dunham, & Pierce, 1998; Marsh & Hocevar, 1988). Results from Model 4 met both of these basic requirements. Campbell and Fiske (1959) also suggested that convergent validity requires that monotrait heteromethod correlation coefficients be significant and substantial in magnitude and be higher than heterotrait monomethod (i.e., method effects) or heterotrait heteromethod coefficients. Because correlations among CFA factors essentially represent correlations among scale scores corrected for attenuation, the Campbell and Fiske criteria of determining the convergent and discriminant validity can be readily and more accurately applied (Marsh & Hocevar, 1988). As Table 3 reports, the convergent validity (i.e., monotrait heteromethod) coefficients were clearly higher than the heterotrait heteromethod coefficients. Convergent validity coefficients were not always higher than the heterotrait monomethod correlation coefficients in the same column or row because some of the self-efficacy factors were highly correlated when they shared the same method. Still, on the whole, the monotrait heteromethod correlation coefficients were generally higher (average $r = .612$) than either the average correlation among Problem Self-Efficacy factors (average $r = .449$) or the average correlation among the MSLQ Self-Efficacy factors (average $r = .392$). Self-efficacy factors for the six subjects also showed a pretty consistent pattern of interrelatedness. Across the two methods, English and History Self-Efficacy and the three math-related self-efficacy factors (i.e., Algebra, Geometry, and Chemistry Self-Efficacy) demonstrated particularly strong correlations.

Although these results are strongly suggestive of convergent validity, the nature of analyses does not permit us to generate precise answers. In particular, it is difficult to separate out the relative contribution of trait and method effects from these trait–method combination first-order factors. Marsh and Hocevar (1988) demonstrated that this could be achieved by applying HCFAs to the MTMM matrix. However, to obtain a fully identified second-order factor structure, a minimum of three trait and three method first-order factors need to be defined.
STUDY 2

Because only two methods were used in Study 1, it was not possible to analyze relations among trait factors after controlling for the method effects, nor was it possible to examine equivalence of different methods after the trait effects had been accounted for. In Study 2, three different methods were used for assessing self-efficacy judgments across three different subject areas. More specifically, students’ self-efficacy perceptions in Korean, English, and math (hence three traits) were assessed by problem-referencing, task-referencing, and MSLQ (i.e., subject-referencing) methods (hence three methods). Therefore, it became possible to examine the aforementioned issues directly by specifying CFA and HCFA models, in addition to interpreting findings more clearly according to the Campbell and Fiske (1959) guidelines.

Method

Participants

Participants were 235 students from a Korean all-female high school. Students completed self-efficacy surveys as part of a research project comparing the predictive utility of different self-efficacy beliefs for immediate and delayed academic performances (Bong, in press).

Measures and Procedures

Problem-referenced self-efficacy. Problems used for assessing self-efficacy perceptions came from placement tests developed by one of the educational testing services in Seoul, Korea. These tests contained problems of representative types and topics that entering high school students should be able to handle. There were 25 problems in each school subject. In the Korean test, several long reading passages were used, each of which was often referred to by multiple problems. Some of these questions also referred to each other, making it difficult to separate them. In these instances, presenting individual problems was not deemed appropriate because doing so would remove them from their very contexts. When these problems were inspected across passages, several common problem types were identified (e.g., vocabulary, grammar, reading comprehension, etc.). Therefore, problems were reorganized according to their types and each reading passage was presented with one or more problems that it most logically related to. This resulted in reduction in the number of problems presented in Korean ($n = 15$). Students rated their confidence toward solving problems of the given type when these problems were presented for a brief duration. Procedures were the same as those used in Study 1.
Task-referenced self-efficacy. Ten representative task descriptions in each domain were developed out of the placement test problems. All specific information such as numbers, figures, vocabulary, or reading passages were removed from problems. These problem descriptions were then revised so that they illustrated generic portraits of tasks typically performed in each domain. For example, students read task descriptions such as “Read a given passage and determine its main theme” and “Change given sentences from active to passive voice” in Korean, “Read a given paragraph and fill in parentheses with appropriate conjunctions” and “Find parts that are grammatically incorrect from given sentences” in English, and “Solve equations containing square roots” and “Solve for x in a quadratic equation” in math. A response scale ranging from 0 to 100 was used again with the same verbal descriptors (see Appendix for task descriptions).

MSLQ self-efficacy. Among the six MSLQ items used in Study 1, only five items were retained in Study 2. The item that read “My study skills are excellent in (a subject) class” was dropped. Self-efficacy refers to personal convictions and expectations and items measuring self-efficacy should thus ask whether one is confident that she “can” or “will be able to” execute certain behaviors required for desired outcomes. In this sense, the particular item in question did not exactly appear to tap perceived efficacy. Although empirical results from Study 1 show that this item behaved similarly to other items, it was removed from the scale for this conceptual reason. A response scale ranged from 1 (not at all true) to 5 (very true) to make it consistent with other response scales used in the survey.

Results

As was the case in Study 1, we present results from the item-level analyses before reporting on the full MTMM models that used item parcels. Again following Bandalos and Finney’s (2001) recommendation, we tried to model the data using individual items as indicators and see whether the items loaded on their intended scales. When model modifications were called for, we examined whether there existed theoretical grounds to specify these effects explicitly in the model. With the problem-specific self-efficacy data, the initial three-factor model with no cross loading and no correlated error did not fit the data well, \( \chi^2(2012, N = 235) = 5,432.019, p < .001 \) (NNFI = .791, CFI = .797). There were 15, 25, and 25 MVs for Korean, English, and Math Problem Self-Efficacy, respectively. Modification indexes suggested numerous correlations among the uniqueness terms, first 70 of which were within the boundary of each scale. Incorporating 63 of such error covariations improved the model fit substantially and to the marginally acceptable level, \( \chi^2(1944, N = 235) = 3,599.409, p < .001 \) (NNFI = .895, CFI = .902). Although the number of correlated error terms was somewhat large, it was not completely unreasonable in light of the large number of items on each scale. We also
found it reassuring that modification indexes primarily suggested correlations between error terms that belonged to the same scale. We inspected the error correlations closely in an attempt to locate some recurring patterns that might point to potential secondary factors in operation that were of substantive or theoretical import. However, a majority of these correlations appeared to be random in nature, sharing variances due to similar wording, similar stems, or some idiosyncrasies in items. By far the most conspicuous pattern was that items that were presented contiguously had errors that were correlated. In fact, 32 of the 63 error covariances incorporated in the final model dealt with contiguous items.

An item-level CFA model was next fitted within the task-specific scales. Fit of the initial three-factor model with 10 MVs each (i.e., Korean, English, and Math Task Self-Efficacy) with no cross loading and no correlated error fell a little short of the acceptable criteria, $\chi^2(403, N = 235) = 1,197.987$, $p < .001$ (NNFI = .857, CFI = .868). Six correlation paths between error terms significantly improved the model fit, $\chi^2(397, N = 235) = 935.803$, $p < .001$ (NNFI = .902, CFI = .910). All correlated error paths were within the same scale, one in Korean, four in English, and one in math. In the case of Korean, the correlated errors were from items that had similar sentence beginnings, “read a given material,” although one referred to a poem and the other to reading passages. Other items with similar beginnings did not share their uniquenesses. In English, error terms again tended to correlate between items that had similar stems (e.g., “read a given passage,” “read given sentences,” or “read a given paragraph”). One correlation appeared to be between items that tapped on some aspects of English grammar. The single correlated error path in math was between items that involved equations. It was also notable that all six correlated error paths were between items that were contiguously placed in the survey, although not all errors of contiguously positioned items were correlated.

Although some of these correlated errors made sense conceptually, the pattern in neither problem-specific nor task-specific scales was deemed strong or consistent enough to require explicit modeling. We thus decided that these effects were not of much theoretical interest in this context. When the item-level model was fitted to the MSLQ data, the initial fit was not satisfactory, $\chi^2(87, N = 235) = 537.170$, $p < .001$ (NNFI = .787, CFI = .824). Modification indexes suggested the need for correlated error paths between items that shared the same wording. Model fit became satisfactory with these modifications, $\chi^2(72, N = 235) = 118.266$, $p < .001$ (NNFI = .974, CFI = .982). In contrast to the error–uniqueness correlations in problem- and task-specific self-efficacy scales, these effects were consistent and substantively meaningful. Therefore, we modeled these effects indirectly in all subsequent models by correlating errors between item parcels that were made up of items with the same wording.

As was the case in Study 1, we compared the factor correlation coefficients from the within-scale, item-level models with those from the full MTMM model based on item parcels. The differences ranged between .007 and .023 (average =
among problem-specific scales, between .005 and .029 (average = .018) among task-specific scales, and between .001 and .012 (average = .008) among MSLQ scales. These results provided us with some assurance that the results would not have been much different had the individual items been used in lieu of item parcels. To minimize confounding, we again used item parcels for all scales.

A total of 39 MVs were created (descriptive statistics of individual items and MVs are available from Mimi Bong). With English and Math Problem Self-Efficacy, responses to Problems 1, 6, 11, 16, and 21 were aggregated to produce the first MV. Following the next sequences produced another 4 MVs. In Korean, combining three responses according to the same sequence (starting with Problems 1, 6, and 11) produced 5 MVs. With Task Self-Efficacy, 5 MVs in each subject were prepared by averaging two responses (starting with Tasks 1 and 6), following similar sequences used in Problem Self-Efficacy. MSLQ responses to Items 1 and 5 and Items 2 and 4 were combined and Item 3 functioned as a single-item indicator. In total, 5 Problem Self-Efficacy MVs, 5 Task Self-Efficacy MVs, and 3 MSLQ Self-Efficacy MVs in each school subject were constructed.

Table 4 presents goodness-of-fit indexes for all CFA and HCFA models. Models 1 to 4 in Study 2 shared the same theoretical structure with Models 1 to 4 in Study 1. The only difference between the two studies with regard to these four models was the number of trait and method factors specified. Results from these first-order CFAs were consistent with those obtained in Study 1. Specifying method factors only in the absence of trait effects (Model 1) or trait factors only without method effects (Model 2) resulted in poor model fit, although the trait-only model was somewhat superior compared to the method-only model as was the case in Study 1. Model 3 that separated out trait and method variance at the item-parcel level by specifying three method and three trait first-order factors demonstrated marginally acceptable model fit (NNFI = .885, CFI = .900). Model 4 with nine first-order factors, each of which reflected unique combination of a single method and a single trait, demonstrated the best and satisfactory fit to the empirical data (NNFI = .949, CFI = .955). Standardized factor loadings ranged from .668 to .972 (Mdn = .912).

Examining correlation coefficients among these nine factors again permitted applying the Campbell and Fiske (1959) rules for determining convergent and discriminant validity. As Table 5 shows, factor intercorrelation in Model 4 provides evidence of convergent validity of traits assessed by different methods as well as discriminant validity of traits assessed by the same method. More specifically, monotrait heteromethod coefficients between Problem and Task Self-Efficacy factors were .680 in Korean, .790 in English, and .712 in math, with an average correlation of .727. The three correlation coefficients were higher than all heterotrait monomethod and heterotrait heteromethod coefficients in the same column or row. Convergent validity coefficients between Task and MSLQ Self-Efficacy factors came next in magnitude, ranging from .600 to .751 (average r = .669).
Those between Problem and MSLQ Self-Efficacy factors were .611 in Korean, .593 in English, and .563 in math, with an average correlation of .589. These coefficients were clearly larger than the heterotrait heteromethod coefficients. With few exceptions, the monotrait heteromethod correlation coefficients were also larger than heterotrait monomethod coefficients. The difference, however, was not definitive. Campbell and Fiske argued that such a case represents true trait correlation, strong method effects, or both. All correlation coefficients among different self-efficacy factors that shared the same method were well below 1.0 across the three methods, attesting to the discriminant validity of the three subject-specific self-efficacy factors. The Korean, English, and Math Self-Efficacy factors demonstrated an average correlation of .560, .625, and .336 when assessed with the problem-referencing, task-referencing, and MSLQ methods, respectively.

The fundamental difference of the MTMM matrix in Study 2 from that in Study 1 was the provision of three methods. Whereas trait and method effects had to be

### Table 4

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>( \chi^2 )</th>
<th>df</th>
<th>NNFI</th>
<th>CFI</th>
<th>Res.</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>39 uncorrelated first-order factors</td>
<td>12367.610</td>
<td>741</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>3 correlated method first-order factors only</td>
<td>6196.917</td>
<td>690</td>
<td>.491</td>
<td>.526</td>
<td>.106</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>3 correlated trait first-order factors only</td>
<td>3771.702</td>
<td>690</td>
<td>.706</td>
<td>.726</td>
<td>.089</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>3 correlated trait and 3 correlated method first-order factors with no trait–method correlation</td>
<td>1812.744</td>
<td>648</td>
<td>.885</td>
<td>.900</td>
<td>.076</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>9 correlated trait–method combination first-order factors</td>
<td>1180.901</td>
<td>657</td>
<td>.949</td>
<td>.955</td>
<td>.035</td>
<td>1.00</td>
</tr>
<tr>
<td>4a</td>
<td>9 uncorrelated trait–method combination first-order factors</td>
<td>2496.384</td>
<td>693</td>
<td>.834</td>
<td>.845</td>
<td>.343</td>
<td>.000</td>
</tr>
<tr>
<td>A</td>
<td>3 correlated method second-order factors</td>
<td>1654.587</td>
<td>681</td>
<td>.909</td>
<td>.916</td>
<td>.077</td>
<td>.640</td>
</tr>
<tr>
<td>B</td>
<td>3 correlated trait second-order factors</td>
<td>1414.255</td>
<td>681</td>
<td>.931</td>
<td>.937</td>
<td>.061</td>
<td>.822</td>
</tr>
<tr>
<td>C</td>
<td>3 correlated trait and 3 correlated method second-order factors with no trait–method correlation</td>
<td>1186.591</td>
<td>669</td>
<td>.951</td>
<td>.955</td>
<td>.038</td>
<td>.996</td>
</tr>
<tr>
<td>D</td>
<td>Model C with MSLQ method factor removed</td>
<td>1209.047</td>
<td>674</td>
<td>.949</td>
<td>.954</td>
<td>.044</td>
<td>.979</td>
</tr>
</tbody>
</table>

**Note.** NNFI = Bentler–Bonnett nonnormed fit index; CFI = comparative fit index; Res. = average absolute standardized residuals; TC = target coefficient \([\chi^2 \text{ for Model 4a} - \chi^2 \text{ for the model being tested})/(\chi^2 \text{ for Model 4a} - \chi^2 \text{ for Model 4})\]; MSLQ = Motivated Strategies for Learning Questionnaire.
<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean Problem SE</td>
<td>1.000</td>
<td>.638</td>
<td>.506</td>
<td>.790</td>
<td>.660</td>
<td>.460</td>
<td>.611</td>
<td>.257</td>
<td>.214</td>
</tr>
<tr>
<td>English Problem SE</td>
<td>.353</td>
<td>1.000</td>
<td>.571</td>
<td>.442</td>
<td>.642</td>
<td>.712</td>
<td>.310</td>
<td>.751</td>
<td>.563</td>
</tr>
<tr>
<td>Math Problem SE</td>
<td>.469</td>
<td>.653</td>
<td>1.000</td>
<td>.1000</td>
<td>.712</td>
<td>.442</td>
<td>.258</td>
<td>.656</td>
<td>.279</td>
</tr>
<tr>
<td>Korean Task SE</td>
<td>.680</td>
<td>.506</td>
<td>.469</td>
<td>1.000</td>
<td>.587</td>
<td>.1000</td>
<td>.249</td>
<td>.656</td>
<td>.303</td>
</tr>
<tr>
<td>English Task SE</td>
<td>.620</td>
<td>.790</td>
<td>.706</td>
<td>.712</td>
<td>1.000</td>
<td>.249</td>
<td>.460</td>
<td>.354</td>
<td>.303</td>
</tr>
<tr>
<td>Math Task SE</td>
<td>.606</td>
<td>.611</td>
<td>.310</td>
<td>.469</td>
<td>.460</td>
<td>1.000</td>
<td>.249</td>
<td>.354</td>
<td>.303</td>
</tr>
<tr>
<td>Korean MSLQ SE</td>
<td>.611</td>
<td>.593</td>
<td>.310</td>
<td>.460</td>
<td>.600</td>
<td>.1000</td>
<td>.249</td>
<td>.354</td>
<td>.303</td>
</tr>
<tr>
<td>English MSLQ SE</td>
<td>.257</td>
<td>.563</td>
<td>.258</td>
<td>.460</td>
<td>.751</td>
<td>.303</td>
<td>1.000</td>
<td>.249</td>
<td>.354</td>
</tr>
<tr>
<td>Math MSLQ SE</td>
<td>.214</td>
<td>.241</td>
<td>.249</td>
<td>.354</td>
<td>.751</td>
<td>.303</td>
<td>.249</td>
<td>1.000</td>
<td>.303</td>
</tr>
</tbody>
</table>

Note. Bold entries represent monotrait heteromethod coefficients; italic entries represent heterotrait monomethod coefficients. SE = Self-Efficacy; MSLQ = Motivated Strategies for Learning Questionnaire.
directly inferred from the relations among MVs in Study 1, these effects could now be estimated from relations among the nine trait–method combination first-order factors. Second-order trait and second-order method factors could be separately identified on the basis of these first-order factors. As a result, it was possible to examine either trait correlation after method effects were removed or method correlation after trait effects were accounted for. Table 4 presents the goodness-of-fit indexes of the second-order CFA models tested.

Model A specified three correlated second-order method factors only, whereas Model B postulated three correlated trait factors only. If any of these models demonstrated acceptable fit, it would mean that the covariation among the nine trait–method combination factors was primarily created by either the trait or method effects only. In evaluating the fit of HCFA models, relying solely on the usual goodness-of-fit indexes such as NNFI and CFI could be misleading. They only indicate the overall ability of models for depicting the indicator variance. In cases in which first-order factors are relatively uncorrelated, values of NNFI or CFI of second-order models can be quite high even when the second-order factors explain little of the first-order factor variance. Marsh and Hocevar (1985) proposed an index that is sensitive to the degree of first-order factor correlation that can be used in determining the fit of higher order models. The target coefficient (TC) roughly represents the proportion of first-order factor variance that is accounted for by the second-order factors. When used along with traditional fit indexes, it provides more accurate information to researchers about the usefulness of the second-order factor structures.

As shown in Table 4, Model A with second-order method factors only and Model B with second-order trait factors only produced suitable NNFI and CFI values, indicating that the models as a whole were able to account for covariation among the MVs to a reasonable degree. However, the second-order structures were not able to illustrate the first-order factor covariance to a sufficient degree as evidenced by the TC values of .640 and .822, respectively. Model C with both trait and method second-order factors not only demonstrated superior overall fit compared to Model A, $\Delta \chi^2(12, N = 235) = 467.996, p < .001$, or Model B, $\Delta \chi^2(12, N = 235) = 227.664, p < .001$, but also displayed an excellent TC (.996). Most of the covariation among the first-order factors was thus accounted for by the hypothesized trait and method second-order factors.

Table 6 reports correlation coefficients among the second-order factors of Model C. These coefficients are especially helpful in answering the convergent and discriminant validity questions. For example, the correlation coefficients among the second-order self-efficacy factors were obtained after the method effects were accounted for. This allowed us to put more faith in the answers generated from these coefficients than those from the correlation among trait–method combination first-order factors. Korean, English, and Math Self-Efficacy factors appeared sufficiently distinct from each other. Correlation coefficients were of moderate value, ranging from .371 to .460. Because these coefficients were not too
high and substantially different from unity, discriminant validity of these self-efficacy factors was supported. Deciding whether a given correlation coefficient between any two traits is too high to judge them different calls for a rather subjective judgment (see, e.g., Marsh & Hocevar, 1988, for related discussion). In our context, these correlation coefficients should be perceived as true trait correlation discussed by Campbell and Fiske (1959) because they represent relations among trait factors that are corrected for unreliability and independent of the shared method variance. The fact that consistent trait correlation was observed across methods in the first-order CFA (Table 5) also supports true trait correlation. Among the second-order method factors, the Problem-Referencing and Task-Referencing factors were most highly correlated (.688). The Task-Referencing factor was also moderately correlated with the MSLQ factor (.529). There was very weak correlation between the Problem-Referencing and MSLQ factors (.124).

In HCFA, evidence of convergent validity can be found in the substantial loadings of first-order factors on their respective second-order trait factors. All first-order factors of Model C demonstrated standardized factor loadings above .605 on the second-order traits factors. However, the strong method effects of Problem-Referencing factors (.504–.698) and Task-Referencing factors (.557–.760) on their respective lower order factors should qualify this finding. In comparison, trait effects mostly determined variance of the MSLQ first-order factors. It is also interesting to note that magnitude of the higher order trait effects were very similar on the Problem and Task Self-Efficacy first-order factors (.605 and .608 in Korean, .668 and .693 in English, and .676 and .737 in math) but noticeably greater on the MSLQ first-order factors (1.0 in Korean, .805 in English, and .812 in math). In fact, after the trait effects were removed, the MSLQ method shared virtually no variance with the Problem-Referencing method. This appeared to have caused the disturbance term of the MSLQ Korean Self-Efficacy factor to be fixed to zero. Because the MSLQ first-order factors did not seem to have either enough variance to support both trait and method second-order factors or enough shared variance with the other two scales, method effects were not separately modeled from the MSLQ

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean SE</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English SE</td>
<td>.460</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math SE</td>
<td>.371</td>
<td>.395</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem Referencing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task-Referencing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.688</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>MSLQ</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.124</td>
<td>.529</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note. SE = Self-Efficacy; MSLQ = Motivated Strategies for Learning Questionnaire.
factors. This revised HCFA model (Model D) avoided the problem of zero residual variance in the MSLQ Korean Self-Efficacy factor. Path coefficients changed slightly from those of Model C, but the general pattern remained the same. Figure 1 presents standardized path coefficients and residuals of Model D.

DISCUSSION

Developing a specific self-efficacy scale that is both consistent with theoretical provisions and methodologically sound poses a unique challenge. Oftentimes,
items used for assessing problem-specific self-efficacy are the very problems on achievement tests that students are to solve soon afterwards. Constructed in this fashion, investigators typically do not have enough room to maneuver to make the scale better meet the traditional measurement requirements. Bandura (1997) recognized this by stating that

Efficacy beliefs do not share the major properties ascribed to personality traits. This raises questions about the appropriateness of some of the trait-based psychometric procedures for evaluating self-efficacy measures. … Restricting items to those that correlate highly with one another results in a self-efficacy scale that measures redundantly only a segment of perceived efficacy and perhaps a narrow segment at that. (p. 45)

Yet Bandura suggested that factor analytic studies could shed some light on the soundness of self-efficacy scales.

This investigation compared different measures of academic self-efficacy beliefs across varied subject areas and samples, using CFA approaches to the MTMM data. Convergent validity of self-efficacy scores assessed by different methods and discriminant validity of self-efficacy beliefs across multiple academic domains were examined. Results of the first-order CFAs from Studies 1 and 2 show strong convergence, demonstrating the generalizability of findings. Models that excluded either the trait or the method effects were not able to illustrate the self-efficacy data effectively. The need for different self-efficacy factors confirms the context specificity of self-efficacy perceptions. When gauging their academic confidence, students responded differently depending on what subject matter area was being tapped by each question. At the same time, the need for different method factors indicates that their responses differed to some degree depending on how the questions were posed and what kind of questions were asked.

When the Campbell and Fiske (1959) guidelines were applied to the first-order CFA models, results generally support the convergent validity of self-efficacy beliefs. Across Studies 1 and 2, students’ self-efficacy responses in the same domain assessed by different methods were more highly correlated than self-efficacy scores in different domains assessed by either the same or different methods. Results also verify discriminant validity of self-efficacy responses in different school subjects. Regardless of how they were assessed, self-efficacy scores in different academic domains did not correlate too highly to cast doubt on their distinctiveness. Most correlation coefficients among self-efficacy factors within the same method were considerably less than unity, despite the fact that they were corrected for attenuation due to measurement errors and hence tended to be higher than what we usually observe in the literature. Even when they were highly correlated, as were algebra and geometry self-efficacy scores in Study 1, treating them as products of a single self-efficacy construct resulted in substantial decrement in the ability of models to account for the self-efficacy matrix. Given these results, the moderate correlation among self-effi-
cacy responses should be viewed as evidence of true trait correlation rather than lack of discriminant validity (Campbell & Fiske, 1959).

In fact, the moderate correlation among different academic self-efficacy factors found in this study is precisely what the self-efficacy theory would predict (Bandura, 1997). Prior successes and failures in a given domain are the major determinants of people’s self-efficacy perceptions in that very domain. These perceptions do generalize, however, to the extent that individuals realize that different domains require similar subskills, that dissimilar skills in different domains are acquired and developed concurrently, or that success in various domains depends on common self-regulatory capabilities. It is also the case that students’ achievement levels in diverse subject areas are often highly correlated. Therefore, it is only reasonable to expect students’ self-efficacy judgments in multiple academic subjects to be moderately correlated. The strength of confidence to perform successfully in algebra would be more highly correlated with confidence beliefs in geometry than confidence in, for example, English. However, because there are also skills and competence that are unique to algebra or geometry, strengths of confidence students express toward these two math domains would not be completely identical. Results from this investigation are consistent with Bong’s (1997) previous observations with a larger U.S. sample (N = 588) and Korean middle and high school samples (Bong, 2001a).

Although results discussed up to this point are coherent across Studies 1 and 2, consistent with previous reports and reasonable in light of the self-efficacy theory, it should be reminded that they were based on the factors that did not separate out trait effects from method effects or vice versa. In essence, the only major advantage of these first-order CFAs to the traditional zero-order correlation approach is that they account for unreliability in the measures and thus provide more accurate information. They still suffer the same criticisms that the zero-order approach faces. For example, Marsh and Hocevar (1983) wrote, while analyzing the MTMM matrix of nine traits and two methods

Testing the second and third criteria [proposed by Campbell and Fiske, 1959] alone requires that each of the nine convergent validities be compared with 32 different correlations—a total of 288 comparisons. Besides being unwieldy, the likelihood of obtaining rejections due to sampling fluctuations alone increases geometrically with the number of traits and methods. (p. 233)

By applying HCFAs, researchers are exempt from making the numerous comparisons they otherwise have to make with the zero-order or the first-order CFA correlation. Yet the HCFAs provide more definitive answers regarding the convergent and discriminant validity of scores by isolating the observed variability among responses into different sources. In Study 2, we were able to partition the variance in each self-efficacy scale into trait, method, and residual variance by performing
second-order CFAs. This approach allowed us to examine the degree to which self-efficacy beliefs in different academic subjects correlate, considering the method effects that were apparently in operation. Similarly, it permitted us to assess the degree to which different methods converged with each other after the trait effects were accounted for.

Results corroborated findings from the first-order models regarding the discriminant validity of various self-efficacy judgments. After the variance due to methods and uniquenesses were taken out, correlation among self-efficacy beliefs in three subject areas observably decreased and was only moderate in magnitude. When it comes to convergent validity, however, some interesting and also somewhat unsettling results emerged. Students’ self-efficacy judgments as assessed by specific problems converged with those estimated using specific task descriptions. On one hand, this may be fully expected because task descriptions in our study were initially developed from the self-efficacy assessment problems. On the other hand, it gives researchers some reassurance that resorting to generic task descriptions in lieu of particularized problems would produce approximately similar results. For example, investigators have been using task descriptions that are analogous to the ones used in this study in domains in which problem-specific measurement of self-efficacy is not feasible (e.g., Joo et al., 2000; Pajares et al., 1999; Schunk & Ertmer, 1999; Shell et al., 1995). Our findings provide empirical justification for such practice by establishing reasonable equivalence of problem-referenced and task-referenced self-efficacy ratings.

Self-efficacy responses generated by the task-referencing method were also moderately correlated with those from the MSLQ items. However, the strength of the relation between the problem-referencing method and MSLQ was not to the extent that one could view the two methods comparable. After the trait effects were controlled for, students’ self-judged efficacy toward specific problems displayed practically no relation with self-efficacy ratings from the MSLQ scale. The MSLQ Self-Efficacy scale differs from the other two measures in its generality. One may argue, therefore, that any observed difference between the MSLQ and problem- or task-specific self-efficacy responses is due to their difference in measurement specificity. This is indeed a plausible assumption according to the self-efficacy theory. Bandura (1997) and other self-efficacy researchers (Pajares, 1996; Zimmerman, 1995) have discussed that self-efficacy can and should be assessed at different levels of generality depending on the outcomes of interest. Therefore, one could measure self-efficacy for performing a particular task under a very specific set of conditions (most specific levels), completing a class of activities sharing the common conditions and properties within the same domain (intermediate levels of specificity), or functioning successfully in given domains without identifying the tasks and conditions under which these tasks are to be performed (most general levels). The problem- and task-referencing methods are akin to the most specific
and intermediate levels of specificity, whereas the MSLQ scale resembles the most general levels of self-efficacy assessment. Our findings can thus be taken to substantiate these previous observations that evaluation of one’s competence generates related but unequal inferences depending on the specificity of the contexts provided.

Assuming that differences between the scales observed in this study were mainly a consequence of different measurement specificity, there is little doubt that they would all meaningfully relate to some domain-related outcomes. Even so, it is one thing to establish convergent validity of measures and another to establish their predictive utility, for people build and act on not only specific task beliefs but also general beliefs that are more than sum of their specific beliefs (Bong & Clark, 1999). It is intriguing in this sense that, after the trait effects were controlled, the MSLQ Self-Efficacy scale did not share much of its remaining variance with more specific self-efficacy scales. Again, we speculate that difference in the assessment specificity was a factor that contributed to such results. Nevertheless, generality of self-efficacy assessment scales need not be associated with the lack of concrete and explicit anchors, which often form the basis of requisite judgments. Asking students to gauge their confidence for obtaining specific letter grades in the given course (Zimmerman & Bandura, 1994), for example, is a general-level self-efficacy measurement that still offers specific anchors. Future research should determine whether and how general self-efficacy perceptions, formed in either the presence or the absence of concrete anchors, differ from more specific self-efficacy perceptions and from each other.

At minimum, our results highlight the fact that what and how questions are asked makes a difference in self-efficacy assessment. Researchers would be well advised to pay particular attention to the specificity of self-efficacy measures in view of their predictive and explanatory goals. When devising specific self-efficacy measures, they should take heed to what problems and activities to include, given the strong method effects associated with problem- and task-referencing methods. The question ultimately boils down to selecting a measure that best captures individual’s beliefs as they are faced with specific tasks and contingencies (Pajares, 1996). We recommend that future research that aims to evaluate various self-efficacy measures should also compare the predictive validity of those measures. This investigation also confirmed the effectiveness of higher order factor analytic procedures for analyzing the convergent and discriminant validity issues. As demonstrated here, results may not show the whole picture and subsequent conclusions could be deficient or misleading when one only looks at the zero-order or even first-order factor correlation according to the Campbell and Fiske (1959) criteria. We echo with Marsh and Hocevar’s (1988) suggestion that HCFA should be the choice of analysis when researchers are dealing with MTMM matrices involving at least three traits and three methods.
ACKNOWLEDGMENTS

Portions of Study 1 were presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 1998. An earlier version of this article was presented at the annual meeting of the American Educational Research Association, Seattle, WA, April 2001.

REFERENCES


**APPENDIX**

**Task-Specific Self-Efficacy Items**

How confident are you that you can successfully solve or perform each of the following types of problems or tasks?

**Korean**

1. Read given sentences and correctly infer the meaning of underlined words.
2. Read a given passage and determine its main theme.
3. Change given sentences from active to passive voice.
4. List necessary elements in each type of prose.
5. Read multiple passages and categorize them by type.
6. Correctly interpret the underlined part of a given poem.
7. Classify different metaphors in a poem according to their meaning.
8. Explain methods of writing different styles of prose.
9. Comprehend given sentences using Chinese idioms or proverbs.
10. Reorganize given paragraphs according to their correct order.

**English**

1. Correctly interpret underlined words in a given sentence.
2. Read a given paragraph and fill in parentheses with appropriate conjunctions.
3. Find parts that are grammatically incorrect from given sentences.
4. Provide appropriate responses for questions in a given conversation.
5. Read a given passage and determine its main theme.
6. Read a given passage and come up with appropriate title words.
7. Read given sentences and figure out contents that will appear next.
8. Select a correct synonym for given vocabulary.
9. Read a given passage and answer questions that are given in English.
10. Given multiple conversations, select inappropriate or inadequately combined pairs.
Math

1. Solve equations containing square roots.
2. Solve for $x$ in a quadratic equation.
3. Given vertex and/or Cartesian coordinates, solve a function.
4. Compute a mean, standard deviation, and variance using a frequency table.
5. Given three sides of a triangle, determine whether it is an acute or obtuse triangle.
6. Given perimeter of a figure, compute its area.
7. Solve for a particular angle of a figure that is inscribed in a circle.
8. Given sides of a triangle, compute the area of a circle inscribed in it.
9. Compute the length of a line connecting a particular point on a circle and a tangent line.
10. Solve equations containing $\cos$, $\sin$, $\tan$, $\cos^2$, $\sin^2$, or $\tan^2$. 