Coaching Efficacy and Exploratory Structural Equation Modeling: A Substantive-Methodological Synergy

Nicholas D. Myers, Melissa A. Chase, Scott W. Pierce, and Eric Martin

The purpose of this article was to provide a substantive-methodological synergy of potential importance to future research in sport and exercise psychology. The substantive focus was to improve the measurement of coaching efficacy by developing a revised version of the coaching efficacy scale (CES) for head coaches (N = 557) of youth sport teams (CES II-YST). The methodological focus was exploratory structural equation modeling (ESEM), a methodology that integrates the advantages of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) within the general structural equation model (SEM). The synergy was a demonstration of how ESEM (as compared with CFA) may be used, guided by content knowledge, to develop (or confirm) a measurement model for the CES II-YST. A single-group ESEM provided evidence for close model-data fit, while a single-group CFA fit significantly worse than the single-group ESEM and provided evidence for only approximate model-data fit. A multiple-group ESEM provided evidence for partial factorial invariance by coach’s gender.

Keywords: youth sport, factorial invariance, geomin rotation, simulation, categorical data

For several decades factor analysis has been closely linked with investigations of construct validity (e.g., Nunnally, 1978). Investigations of construct validity in exercise and sport (e.g., scale development) have frequently occurred in studies where only factor analytic measurement model(s), exploratory and/or confirmatory, were specified—typically guided by imperfect measurement theory (Myers, Ahn, & Jin, in press). Within such studies, substantive theory is frequently used both in item development and in forming an a priori hypothesis regarding the underlying factor structure.

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National...
Council on Measurement in Education, 1999) includes at least three relevant guidelines for the substantive focus of the current article. First, an instrument (e.g., CES; Feltz, Chase, Moritz, & Sullivan, 1999) should be revised when there is clear evidence for ways to increase the validity of the measures produced by it. Second, an instrument that is meaningfully revised should be labeled as such (e.g., CES II-YST). Third, new validity evidence, including explication of changes to the original framework, should be provided.

Coaching Efficacy: The Substance

Coaching efficacy is the extent to which a coach believes he/she has the capacity to affect the learning and performance of his/her athletes (Feltz et al., 1999). The coaching efficacy construct was developed based on both a multidimensional model of teacher efficacy (Denham & Michael, 1981) and on self-efficacy theory (Bandura, 1997). Like the role of teacher efficacy in models of effective teaching, coaching efficacy occupies a central place in models of coaching effectiveness (Horn, 2002).

Coaching efficacy has been linked with a host of theory-based external variables: coaching behavior (Feltz et al., 1999), team winning percentage (Myers, Vargas-Tonsing, & Feltz, 2005), player improvement (Chase, Feltz, Hayashi, & Hepler, 2005), playing experience (Sullivan, Gee, & Feltz, 2006), imagery (Short, Smiley, & Ross-Stewart, 2005), leadership style (Sullivan & Kent, 2003), and team efficacy (Vargas-Tonsing, Warners, & Feltz, 2003). While there is compelling evidence that measures derived from the CES relate with theoretically relevant external variables, there also is substantial evidence that model-data fit for the CES does not meet heuristic values for close fit (e.g., Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). For example, Feltz, Hepler, Roman, and Paiment (2009) reported that across four relevant studies (e.g., Lee, Malete, & Feltz, 2002), fit indices ranged from .87 to .89 (CFI) and .080 to .085 (RMSEA) for the multidimensional measurement model of the CES.

Myers, Wolfe, and Feltz (2005) provided an investigation of the psychometric properties of measures derived from the CES. The authors concluded that (a) there were problems with the rating scale categorization structure, (b) there was limited discriminant validity among game strategy efficacy and technique efficacy, (c) the operational definition for each dimension should be reconsidered, (d) several items needed to be revised and/or dropped, and (e) the resultant measures were relatively imprecise. The authors acknowledged that measures produced by the CES have consistently related to theoretically relevant external variables: “That rather imprecise measures produce evidence of external validity probably speaks to the robustness of the coaching efficacy model. Although there is no substitute for a good theory, modifications are available to increase the precision of measures” (p. 156). A subsequent study provided evidence for noninvariance of some of the CES items by coach’s gender (Myers, Wolfe, Feltz, & Penfield, 2006).

Myers, Feltz, Chase, Reckase, and Hancock (2008) put forth the Coaching Efficacy Scale II—High School Teams (CES II-HST) by revising the CES in accord with the extant relevant research. Two substantive revisions implemented in CES II-HST are important for the current study. First, the intended population was delimited to head coaches of high school team sports where a head coach has the opportunity to meaningfully intervene during competition. This sharpened focus
was consistent with empirical evidence for noninvariance by level coached (high school versus youth sport; Penfield, Myers, & Wolfe, 2008) and the type of sport implied by some CES items (e.g., build team cohesion).

The second substantive revision of importance to the current study is a set of changes to the CES itself. These changes include a new dimension of coaching efficacy, revised operational definitions for two of the previous dimensions, and a majority of new or revised items as compared with CES items. Physical conditioning efficacy is the dimension that is new in CES II-HST. Ability to physically condition athletes is a key expected coaching competency domain (National Association for Sport and Physical Education, 2006). The operational definition for character building efficacy and for technique efficacy were each revised to improve the clarity of the definition, and to decrease the likelihood for empirical redundancy with game strategy efficacy, respectively. In accord with the previous changes, all but 1 of the 18 items that define the CES II-HST either were revised (9 items) or were new (8 items) as compared with the CES.

The multidimensional latent variable model for the CES II-HST posits that five dimensions of coaching efficacy covary and influence responses to the items. Motivation efficacy (ME) is measured by four items and is defined as the confidence coaches have in their ability to affect the psychological mood and psychological skills of their athletes. Game strategy efficacy (GSE) is measured by four items and is defined as the confidence a coach has in his/her ability to lead during competition. Technique efficacy (TE) is measured by four items and is defined as the confidence a coach has in his/her ability to use her/his instructional and diagnostic skills during practices. Character building efficacy (CBE) is measured by three items and is defined as the confidence a coach has in his/her ability to positively influence the character development of her/his athletes through sport. Physical conditioning efficacy (PCE) is measured by three items and is defined as the confidence a coach has in his/her ability to prepare her/his athletes physically for participation in her/his sport.

Validity evidence for measures derived from the CES II-HST is accumulating. Myers, Feltz, Chase, et al. (2008) provided evidence for close model-data fit of the measurement model under a CFA approach: $\chi^2(70, N = 549) = 114, p < .001$, RMSEA = .031, SRMR = .034, CFI = .990, and NNFI = .996. Correlations among latent variables ranged from .40 (TE with CBE) to .78 (TE with GSE). Latent construct reliability ranged from .82 (ME) to .92 (TE). Evidence for factorial invariance by coach’s gender was provided. Myers, Feltz, and Chase (2011) provided evidence for the ability of measures derived from the CES II-HST to be predicted by theoretically relevant sources of efficacy information (e.g., coaching education) and concluded that “. . . the CES II-HST should be viewed as a viable replacement for the CES when the intended population fits the specifications of the revised instrument” (p. 86). Myers, Feltz, Chase, et al. (2008), however, cautioned that “the CES [not the CES II-HST] should be used with these [youth sport] coaches until a suitable replacement is available” (p. 1074).

Head coaches of youth sport are typically unpaid volunteers (Martens & Gould, 1979) who have a unique opportunity to play a key role in the psychosocial development of young athletes (Petitpas, Cornelius, Van Raalte, & Jones, 2005; Smith, Smoll, & Curtis, 1978). Effective coaching at the youth sport level frequently manifests in different ways (e.g., fun and fundamental skill development) as compared with effective coaching manifestations (e.g., winning) at higher
levels of competition (McCallister, Blinde, & Weiss, 2000). Because youth sport coaching is important and is culturally unique, improving the measurement of key constructs at the youth sport level is a vital area of research. Unfortunately, “despite its potential influence on the sport experience of children, the coaching efficacy of coaches at the volunteer youth sport levels has received little attention” (Feltz et al., 2009, p. 25).

There is initial evidence that coaching efficacy is an important construct for effective volunteer leadership in youth sport (Feltz et al., 2009). Feltz et al. provided mixed evidence for model-data fit of the CES measurement model under a CFA approach: $\chi^2(246, N = 479) = 901$, $p < .001$, RMSEA = .08, CFI = .97, and NNFI = .97. Correlations among observed subscale composite scores ranged from .40 (TE with CBE) to .76 (TE with GSE). Feltz et al. attributed indications of close (and improved as compared with previous studies with the CES) fit (e.g., CFI) to a condensed rating scale categorization structure and noted some familiar (as compared with previous studies with the CES) indications of only marginal fit (e.g., RMSEA). The substantive focus of the current article, therefore, was to further improve the measurement of coaching efficacy by developing the CES II-YST, based in part, on advances put forth in the CES II-HST.

**From EFA to CFA to ESEM: The Methodology**

The methodological focus of the current article was ESEM, a methodology that integrates the advantages of EFA and CFA within the general SEM (Asparouhov & Muthén 2009). Over a century ago Spearman (1904) articulated what is now known as EFA. Exploratory factor analysis has since become a widely used multivariate analytic framework in many disciplines. Over the past few decades, however, some limitations in the way EFA is typically implemented in software has likely impeded the use of the technique in favor of CFA—even when a priori measurement theory was insufficient to warrant a confirmatory approach. Limitations in the way that EFA is often implemented in software include the absence of standard errors for parameter estimates, restrictions on the ability to incorporate a priori content knowledge into the measurement model, an inability to fully test factorial invariance, and an inability to simultaneously estimate the measurement model within a broader SEM.

The most common measurement model within SEM is Jöreskog’s (1969) CFA (Asparouhov & Muthén 2009). Confirmatory factor analysis provides standard errors for parameter estimates, allows a priori content knowledge to guide model specification, provides a rigorous framework to test factorial invariance, and allows the measurement model to be a part of a broader SEM. A necessary condition for appropriate use of CFA, however, is sufficient a priori measurement theory. Absence of such theory often results in an extensive post hoc exploratory approach guided by modification indexes. Such an approach is susceptible to producing an accepted model that is inconsistent with the true model—despite possible consistency with a particular dataset (MacCallum, Roznowski, & Necowitz, 1992). Exploratory factor analysis, not CFA, may often be the better framework for post hoc explorations in search of a well-fitting measurement model (Browne, 2001).

A common misspecification in CFA is made when a perfect simple structure is imposed on data that have a more complex structure (Browne, 2001). A perfect
simple structure can be thought of as each observed variable indicating (i.e., “loading on”) only one factor (i.e., variable complexity, \(vc = 1\)). A more complex structure can be thought of as having at least one observed variable indicating (i.e., “cross-loading on”) more than one factor (\(vc \geq 2\)). Erroneously imposing perfect simple structure on complex data often results in upwardly biased covariances between the latent variables and biased estimates in the (non)measurement part of an SEM (Asparouhov & Muthén, 2009; Kaplan, 1988). Both of these problems probably have impeded research in sport and exercise psychology, as well as research with psychological instruments in general (Marsh et al., 2009). For example, the high correlation (~.80) typically observed between TE and GSE (for both the CES and the CES II-HST) may be somewhat inflated owing to misspecification in an overly restrictive (i.e., perfect simple structure) CFA approach.

In Mplus (Muthén & Muthén, 1998–2010), ESEM integrates the relative advantages of EFA and CFA within the general SEM. Like EFA, ESEM imposes fewer restrictions on the measurement model (e.g., allows for complex structure) than common implementations of CFA (e.g., imposing perfect simple structure) and integrates advances in possible direct rotations of the pattern matrix (Jennrich, 2007). Like CFA, ESEM can be part of a SEM, which affords much greater modeling flexibility than observed in the traditional EFA framework. In fact, within a single SEM specification, both CFA and ESEM can be imposed simultaneously. When sufficient a priori measurement theory exists, the more parsimonious CFA is preferred (Asparouhov & Muthén, 2009).

Asparouhov and Muthén (2009) introduced ESEM to the methodological community. The general ESEM is summarized below for continuous outcomes consistent with Asparouhov and Muthén. While the substantive and the methodological part of the current article were both focused primarily on measurement, the fuller statistical model is presented as a broader methodological framework within which some of the next generation of sport and exercise psychology research may occur. The range of parameter estimates typically available in SEM is available in ESEM. The first equation is the measurement model (Bollen, 1989):

\[
Y_{p \times 1} = \nu_{p \times 1} + \Lambda_{p \times m} \eta_{m \times 1} + K_{p \times q} x_{q \times 1} + \epsilon_{p \times 1} \tag{1}
\]

where

- \(p\) = the number of continuous observed dependent variables
- \(m\) = the number of continuous latent variables
- \(q\) = the number of observed independent variables

The second equation is the latent variable model (Bollen):

\[
\eta_{m \times 1} = \alpha_{m \times 1} + \beta_{m \times q} \eta_{q \times 1} + \Gamma_{m \times q} X_{q \times 1} + \zeta_{m \times 1} \tag{2}
\]

The appendix to this article provides a detailed accounting of each array in Equations 1 and 2.

Given that the substantive and the methodological part of the current article were both focused primarily on measurement, EFA (with means) only within the more general ESEM framework is presented below. Note that some arrays in Equation 1 and Equation 2 drop out in Equation 3 and Equation 4. The reduced measurement equation can be written as follows.
\[ Y_{pxl} = \nu_{pxl} + \Lambda_{pem} \eta_{mxl} + \varepsilon_{pxl} \]  

The reduced latent variable model can be written as

\[ \eta_{mxl} = \alpha_{mxl} + \zeta_{mxl} \]

Parameter estimates derived from fitting Equation 3 and Equation 4 to a particular dataset can be used to determine sufficient sample size for a desired level of power using Monte Carlo methods in Mplus (Asparouhov & Muthén, 2009).

**ESEM (as Compared with CFA) With the CES II-YST: The Synergy**

The substantive-methodological synergy (Marsh & Hau, 2007) of the current article was a demonstration of how ESEM (as compared with CFA) may be used, guided by content knowledge, to develop (or confirm) a measurement model for the CES II-YST. Four research questions were posed.

**Research question 1.** How many factors were warranted to explain responses to the CES II-YST? As depicted in Figure 1 and Figure 2 the a priori hypothesis was five factors.

**Research question 2.** Could a more parsimonious/restrictive CFA, informed by the a priori measurement theory, offer a viable alternative to a more complex/flexible ESEM (i.e., compare the parametric simplicity in Figure 1 to the parametric complexity in Figure 2)? The a priori hypothesis was that the CFA would fit significantly worse than the ESEM because initial investigations by definition are typically exploratory to some degree.

**Research question 3.** What was the minimum necessary \( N \) for a desired level of power with regard to the measurement model for the CES II-YST? An a priori hypothesis was not put forth because the development of CES II-YST was in progress (e.g., no previous empirical data were available).

**Research question 4.** Was there evidence for factorial invariance by coach’s gender? An a priori hypothesis was not put forth because the development of CES II-YST was in progress.

**Methods**

**Development of the CES II-YST**

The CES II-HST served as a starting point for the CES II-YST. This approach was considered reasonable given the evidence for the validity framework of the CES II-HST and the fact that the CES II-HST incorporated key tenets of the original validity framework for the CES. Because the validity framework for the CES and the CES II-HST have been detailed elsewhere (e.g., Feltz et al., 1999; Myers, Feltz, Chase et al., 2008), only core aspects of CES II-YST will be described.

The development of the CES II-YST was an iterative process guided by a group of content experts (one faculty member and four graduate students in sport psychology) who critically reviewed the relevant literature (e.g., Côté & Gilbert,
Figure 1 — A priori measurement theory for the CES II-YST from a CFA perspective.
Figure 2 — A priori measurement theory for the CES II-YST from an ESEM perspective. Error terms of latent response variates are omitted from the figure to reduce clutter.
interviewed six youth sport coaches, and consulted with five sport coaching experts. The CES II-YST was developed for head coaches of youth sport teams where a head coach has the opportunity to meaningfully intervene during competition. Age of the athletes was delimited to between 8 and 13 years. The lower age boundary was selected to increase the likelihood of sampling from at least a somewhat competitive sport environment. The upper age boundary was selected because the CES II-HST is intended for the high school level.

The basic structure of the multidimensional measurement model for the CES II-HST was determined to be appropriate for the CES II-YST. The operational definition for each latent variable previously described in the Introduction (ME, GSE, TE, CBE, and PCE) was adopted without revision in the CES II-YST. The item stem from the CES II-HST, “in relation to the team that you are currently coaching, how confident are you in your ability to,” was adopted without revision in the CES II-YST. The rating scale categorization structure in the CES II-YST (a four-category rating scale structure denoting low, moderate, high, and complete confidence) was the same as for the CES II-HST based on the results of Myers, Feltz, and Wolfe (2008).

Item development and the forming of an a priori hypothesis regarding the underlying factor structure of the CES II-YST were both based on a feedback loop between substantive theory and insight provided by content experts. Several items from the CES II-HST (e.g., prepare an appropriate plan for your athletes’ off-season physical conditioning) were determined to be inappropriate for the CES II-YST. Several new items for the CES II-YST were developed (e.g., teach your athletes new skills in a safe manner during practices). Table 1 provides the text for all CES II-YST items. In accord with the cultural uniqueness of youth sport coaching, 9 of the 18 items that define the CES II-YST either were significantly revised or were new as compared with the CES II-HST.

Consistent with the CES II-HST, it was determined a priori that each dimension of the CES II-YST would be defined by only a few items. There were multiple justifications for this decision. Practically, recruiting head coaches for data collection during an athletic season can be an extremely difficult task and one that is made more difficult with an unnecessarily lengthy questionnaire. Conceptually, the authors believe that a few high-quality items can provide sufficient content coverage for each well-defined dimension of coaching efficacy. Empirically, items from both the CES and the CES II-HST have a history of at least a moderately high pattern coefficient value on the intended factor (e.g., Feltz et al., 1999; Myers, Feltz, Chase et al., 2008).

The a priori measurement theory for the CES II-YST from a CFA perspective is depicted in Figure 1. The covariance between the residual variance of me3* and the residual variance of me4* was to be freely estimated because the pair of items was hypothesized to have something in common (i.e., confidence of athletes) in addition to motivation efficacy (see Table 1). As is common in CFA, the model in Figure 1 assumes a perfect simple structure (i.e., \( \nu_c = 1 \) for each variable). The a priori measurement theory for the CES II-YST from an ESEM perspective is depicted in Figure 2. As is necessary in ESEM, the model in Figure 2 allows for a complex structure (i.e., \( \nu_c \) is generally free to \( = m \), 5 in this case, for each variable).
Table 1  Operational Definitions and Items for the Coaching Efficacy Scale II—Youth Sports Teams (CES II-YST)

Motivation Efficacy (ME): confidence a coach has in his/her ability to affect the psychological mood and skills of her/his athletes

*me1: motivate your athletes to work hard
*me2: motivate your athletes to like to participate in sport
me3: help your athletes to maintain confidence in their ability to perform when they are performing poorly
*me4: help your athletes to be confident in their ability relative to their skill level

Game Strategy Efficacy (GSE): confidence a coach has in his/her ability to lead during competition

gse1: make effective strategic decisions in pressure situations during competition
*gse2: develop effective strategies during competition that your athletes understand
gse3: devise strategies that minimize an opposing team’s strengths during competition
gse4: devise strategies that maximize the positive effects of your team’s strengths during competition

Technique Efficacy (TE): confidence a coach has in his/her ability to use her/his instructional and diagnostic skills during practices

te1: instruct all of the different positional groups of your athletes on appropriate technique during practices
te2: teach your athletes the complex technical skills of your sport relative to their skill level during practice
*te3: make corrections for technique errors by playing during practices
*te4: teach your athletes new skills in a safe manner during practices

Character Building Efficacy (CBE): confidence a coach has in his/her ability to positively influence the character development of her/his athletes through sport

cbe1: effectively promote good sportsmanship in your athletes
*cbe2: positively influence a sense of fair play in your athletes
cbe3: positively influence the character development of your athletes
*cbe4: teach life lessons to your athletes through sport

Physical Conditioning Efficacy (PCE): confidence a coach has in his/her ability to prepare her/his athletes physically for participation in her/his sport

*pce1: prepare your athletes to be in physical condition to play the game
pce2: accurately assess your athletes’ physical conditioning

Note. *Denotes that an item was significantly revised or was new as compared with the CES II-HST.

Data Collection

Procedure. An institutional review board provided necessary permission. Data were collected from head coaches of organized youth sport soccer leagues in the United States. Age of athletes ranged from 8 to 13 years. Athletic associations
and directors of coaching assisted with the distribution of the survey to their head coaches (but did not have any access to the data). Data were collected electronically to provide access (at a minimal expense) to a large, national sample of coaches. Informed consent was obtained from all participants. Coaches were assured of confidentiality for their responses. Participants completed the questionnaire no earlier than the second half of a competitive season to ensure adequate experience to make informed judgments.

**Participants.** Total sample size was 557. Head coaches of girls’ teams \(n = 217\), boys’ teams \(n = 155\), and girls’ and boys’ teams \(n = 185\) were represented. A majority of participants identified themselves as Caucasian \(n = 466\) with no other response option endorsed for at least 5% of participants. A majority of participants identified themselves as male \(n = 481; n_{female} = 76\). Age of coach ranged from 19 to 76 years \((M = 42.93, SD = 7.88)\).

**Statistical Modeling**

**Ordinal Data.** Figure 1 and Figure 2 both imply categorical variable methodology (CVM; Muthén, 1984), which is consistent with Myers, Feltz, Chase, et al. (2008). The CES II-YST data were modeled as ordinal under weighted least squares mean-and variance-adjusted estimation (WLSMV). The ordered four-category rating scale structure in the CES II-YST produce data that are non-normal by definition owing to the discrete nature of the metric (Muthén).

Normal theory estimators (i.e., those most commonly employed in SEM) assume that the data follow a conditional multivariate normal (MVN) distribution in the population. As reviewed by Finney and DiStefano (2006), violating the assumption of MVN with categorical data can produce untrustworthy results (e.g., inflated indices of model-data misfit, negatively biased parameter estimates, and negatively biased standard errors). The probability of observing untrustworthy results when categorical data are modeled with a normal theory estimator depends strongly on the degree of non-normality and the number of ordered response options (e.g., DiStefano, 2002; Dolan, 1994; Muthén & Kaplan, 1985). In cases where the number of response options is less than five, Finney and DiStefano suggest using CVM with WLSMV estimation.3

**Rotation.** Rotation of the pattern matrix in ESEM is accomplished via postmultiplication of the pattern matrix by the inverse of an optimal transformation matrix:

\[
\Lambda_{pom} = \Lambda_{pom} (H^*_{mxm})^{-1}
\]  

(5)

An optimal transformation matrix, \(H^*\), is determined by minimizing a continuous complexity function of the elements in the pattern matrix, \(f(\Lambda)\). Geomin (Yates, 1987) was the rotation criterion of interest in the current study. The \(f(\Lambda)\) for geomin implemented in Mplus is

\[
\sum_{i=1}^{p} \left( \prod_{j=1}^{m} \left( \lambda_{ij}^2 + \varepsilon \right) \right)^{1/m}
\]

(6)

where \(\varepsilon\) is a small positive constant added by Browne (2001) to reduce the problem of indeterminancency.
Geomin was selected because it is the default rotation criterion in Mplus and because Asparouhov and Muthén (2009) provided evidence that it performs well when little is known a priori about the true pattern matrix and \( \nu_C \) is likely no more than moderate (i.e., \( \nu_C \leq 3 \)). Oblique rotation was selected to reflect the belief that the dimensions of coaching efficacy covary.

**Rotation Identification.** Some known conditions for rotation identification in factor analysis exist (e.g., Algina, 1980; Hayashi & Marcoulides, 2006). Under oblique rotation \( \mathbf{H}^* \) is a nonsymmetric square matrix that results in \( m^2 \) indeterminacies. Imposing \( m^2 \) constraints on \( \mathbf{\Lambda} \) and \( \mathbf{\Psi} \) is a necessary condition for rotation identification. Fixing the variance for each of the common factors to 1 provided \( m \) constraints. A set of sufficient conditions for imposing the remaining \( m(m-1) \) constraints included: (a) each column of \( \mathbf{\Lambda} \) has \( m-1 \) elements zeros, and (b) each submatrix \( \mathbf{\Lambda}_s \), where \( s=1, \ldots, m \), of \( \mathbf{\Lambda} \) composed of the rows of \( \mathbf{\Lambda} \) that have zeros in the \( m \)th column must have rank \( m-1 \). Consistent with Asparouhov and Muthén (2009), small nonstatistically significant values were counted as zero.

**Model-Data Fit and Reliability.** Indexes of model-data fit considered were \( \chi^2_R \), RMSEA, CFI, and TLI. The standardized root mean squared residual (SRMR) is unavailable under WLSMV estimation. Heuristic classifications model-data fit (e.g., close etc.) were consistent with Hu and Bentler (1999). Construct reliability was measured with coefficient \( H \) (Hancock & Mueller, 2001). Coefficient \( H \) estimates the stability of a modeled construct. Hancock and Mueller suggested that \( H \geq .80 \) was desirable.

**Research Question 1.** The first research question was answered in two steps. Step 1, *number of factors (m)*, considered the fit of a particular ESEM. Step 2, *m - 1 versus m*, considered the relative fit of a simpler ESEM (e.g., unidimensional) as compared with a more complex alternative ESEM (two-dimensional). Six sequential models were fit by systematically increasing the number of factors: Model 1 (unidimensional) to Model 6 (six-dimensional). Nested models were compared with the change in robust \( \chi^2 \) test (\( \Delta \chi^2_R \)). The approach taken in Step 2 is susceptible to over-factoring and inflated Type I error (Hayashi, Bentler, & Yuan, 2007). Therefore \( \alpha \) was set to .01 for these comparisons (\( \alpha = .05 \) otherwise) and the interpretability of the estimated rotated pattern matrix \( \hat{\mathbf{\Lambda}} \) was also considered (i.e., incorporating content knowledge) when deciding which model to accept.

Further, and consistent with weaknesses of the \( \Delta \chi^2_R \) with real data and imperfect theories (Yuan & Bentler, 2004) and the utility of strict adherence to the null hypothesis testing framework with regard to the assessment of model-data fit in general (Marsh et al., 2004), a collection of rough guidelines were also used to judge the magnitude of change in model-data fit for nested models. Consistent with Marsh et al. (2010), \( \Delta \text{CFI} \leq -.01 \) (Chen, 2007; Cheung & Rensvold, 2001), \( \Delta \text{TLI} \leq .00 \) (Marsh, 2007), and \( \Delta \text{RMSEA} \geq .015 \) (Chen), was interpreted as evidence in favor of the more complex model. From this point forward, \( \Delta \text{CFI} \), \( \Delta \text{TLI} \), and \( \Delta \text{RMSEA} \), were collectively referred to as rough guidelines for a nested model comparison.

**Research Question 2.** The second research question was also answered via consideration of \( \Delta \chi^2_R \), \( \Delta \text{CFI} \), \( \Delta \text{TLI} \), and \( \Delta \text{RMSEA} \) because the parametrically simpler CFA in Figure 1 (i.e., Model 8) was nested within the parametrically more
complex ESEM displayed in Figure 2 (i.e., Model 7). As displayed in Figure 2, ESEM imposed fewer restrictions on $A$ than the CFA displayed in Figure 1.

**Research Question 3.** Minimum necessary $N$ for a desired level of power was determined after accepting a measurement model for responses to the CES II-YST. Sample size was determined for a set of parameters of primary interest ($\theta$, where $\theta_i$ was a particular parameter of interest) consistent with Hancock (2006). Monte Carlo methods were used to determine the minimum $N$ at which each $H_0: \theta_i = 0$ was rejected in at least 80% of the replications ($\alpha = .05$) consistent with Muthén and Muthén (2002). Number of replications was set to 10,000 in each run.

**Research Question 4.** Under the theta parameterization, data were fit to four increasingly restrictive multigroup models (ESEM or CFA depending on results from previous research questions) for ordinal data (Millsap & Yun-Tien, 2004). Identification constraints included partial invariance of the thresholds across groups, latent factor means fixed to zero for the reference group (i.e., males), the unique factor covariance matrix fixed to an identity matrix (except for the covariance between me3* and me4*) for the reference group, and the latent intercepts fixed to 0.00 for both groups. The subset of invariant thresholds consisted of the first threshold for all latent response variates and the first and second thresholds for the first latent response variate for each factor. Before testing for factorial invariance, the model was imposed separately in each group, yielding Model 9a (males) and Model 9b (females).

Model 10 through Model 13 tested for factorial invariance by coach’s gender. Model 10 imposed constraints necessary for identification only. Model 11 added the constraint of invariant pattern coefficients to Model 10. Model 12 added the constraint of invariant thresholds to Model 11. Model 13 added the constraint of an invariant residual covariance matrix to Model 12. Nested models were compared via consideration of $\Delta \chi^2_r$, $\Delta$CFI, $\Delta$TLI, and $\Delta$RMSEA. If the $\Delta \chi^2_r$ was statistically significant (and the rough guidelines for nested model comparisons were meaningfully large) then two additional pieces of information were considered: the conceptual ramifications of implementing a particular change based on modification indices (MI) and the $\Delta \chi^2_r$ when parameters of interest were constrained to equality across groups and compared with the previous more complex model.

**Post Hoc Research Questions**

From a mathematical perspective, “measurement invariance does not require invariance in the factor means or factor covariance matrices” (Millsap & Yun-Tien, 2004, p. 484). From a conceptual perspective, however, testing the invariance of these parameters can be informative with regard to substantive theory (Marsh et al., 2009). There has been a lack of clarity with regard to the potential role of coach’s gender (e.g., mean differences) within the conceptual model of coaching efficacy, which has prompted calls for research in this area (with a particular focus on female coaches) while holding level coached constant (Feltz, Short, & Sullivan, 2008).

**Research Question 5.** Was there evidence for latent mean differences on the dimensions of coaching efficacy by coach’s gender? An a priori hypothesis was not put forth because the authors of the current article were unaware of any previous research comparing latent mean differences on dimensions of coaching efficacy.
by coach’s gender at the youth sport level. Model 14 included the restrictions of Model 13 and added the constraint of invariant latent means by coach’s gender.

**Research Question 6.** Was there evidence for invariance of the factor covariance matrix by coach’s gender? An a priori hypothesis was not put forth because the authors of the current article were unaware of any previous research comparing the factor covariance matrix by coach’s gender. Model 15 included the restrictions of Model 14 and added the constrain of an invariant factor covariance matrix by coach’s gender.5

**Results**

**Research Question 1**

The null hypothesis for exact fit was rejected through six-factor extraction (see Table 2). The sixth factor solution was judged not interpretable and likely the result of over-factoring as only one element in the sixth column of \( \hat{\Lambda} \) \( (\hat{\lambda}_{cbe2,6}) \) was statistically significant. The five factor solution (i.e., Model 5) was accepted despite exhibiting statistically significant worse fit than the sixth factor solution (i.e., Model 6), \( \Delta \chi^2_R (13) = 26, \ p = .015 \). This decision was consistent with rough guidelines for a nested model comparison which generally failed to provide evidence in favor of Model 6 as compared with Model 5: \( \Delta \text{CFI} = -.001, \Delta \text{TLI} = -.001, \Delta \text{RMSEA} = .002 \). There was evidence for close (though inexact) fit of Model 5: \( \chi^2_R (72) = 115, \ p < .001, \text{RMSEA} = .033, \text{CFI} = .997, \text{and TLI} = .993. \)

Elements within \( \hat{\Lambda} \) from Model 5 were generally consistent with a priori expectations (compare Table 3 to Figure 1). Note that 12 of the 70 rotated pattern coefficients that were estimated in the ESEM approach, and inconsistent with Figure 1 (e.g., \( \hat{\lambda}_{gse1,ME} \)), were statistically significant (standardized values ranged from 0.15 to 0.51). As depicted in Table 4, interfactor correlations ranged from \( \hat{\psi}_{\text{GSE,CBE}} = .52 \) to \( \hat{\psi}_{\text{GSE,TE}} = .71 (M = .48, SD = .13) \), while construct reliability estimates ranged from \( H_{\text{PCE}} = .89 \) to \( H_{\text{TE}} = .95 \). A reasonable answer to Research Question 1 was that five ESEM factors explained responses to the CES II-YST.

**Research Question 2**

The CFA depicted in Figure 1 (i.e., Model 8) exhibited statistically significant worse fit than the accepted ESEM (i.e., Model 7), \( \Delta \chi^2_R (52) = 335, \ p < .001 \) (see Table 2). Rough guidelines for a nested model comparison also provided evidence in favor of Model 7 as compared with Model 8: \( \Delta \text{CFI} = -.028, \Delta \text{TLI} = -.031, \Delta \text{RMSEA} = .042 \). Further, Model 8 exhibited only approximate fit to the data: \( \chi^2_R (124) = 515, \ p < .001, \text{RMSEA} = .075, \text{CFI} = .969, \text{and TLI} = .962 \). Nonzero elements within \( \Lambda \) from Model 8, however, were consistent with a priori expectations (compare Table 5 to Figure 1). As depicted in Table 4, interfactor correlations from Model 8 ranged from \( \hat{\psi}_{\text{TE,CBE}} = .59 \) to \( \hat{\psi}_{\text{ME,CBE}} = .88 (M = .75, SD = .10) \), while construct reliability estimates ranged from \( H_{\text{PCE}} = .75 \) to \( H_{\text{TE}} = .92 \). The difference between interfactor correlations from Model 8 as compared with Model 7 was positive in all cases (\( M = .27, SD = .08 \)).

Clearly there was evidence that the perfect simple structure imposed in Model 8 may have been erroneous (i.e., some pattern coefficients that were fixed to zero
Table 2  Deriving Measures from the CES II-YST: Key Results from Research Questions 1 and 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Question 1: Number of Factors (m)</th>
<th>Question 1: m – 1 Versus m</th>
<th>Question 2: CFA Versus ESEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2(df)$</td>
<td>$p$</td>
<td>Par</td>
</tr>
<tr>
<td>1: $m = 1$, ESEM</td>
<td>1371(134)</td>
<td>&lt;.001</td>
<td>72</td>
</tr>
<tr>
<td>2: $m = 2$, ESEM</td>
<td>568(117)</td>
<td>&lt;.001</td>
<td>89</td>
</tr>
<tr>
<td>3: $m = 3$, ESEM</td>
<td>316(101)</td>
<td>&lt;.001</td>
<td>105</td>
</tr>
<tr>
<td>4: $m = 4$, ESEM</td>
<td>203(86)</td>
<td>&lt;.001</td>
<td>120</td>
</tr>
<tr>
<td>5: $m = 5$, ESEM</td>
<td>115(72)</td>
<td>&lt;.001</td>
<td>134</td>
</tr>
<tr>
<td>6: $m = 6$, ESEM</td>
<td>90(59)</td>
<td>.006</td>
<td>147</td>
</tr>
</tbody>
</table>

Note. Par = number of parameters estimated; complex = more complex model that a nested model was compared with.
Table 3  Geomin-Rotated Pattern Coefficients (\(\hat{\lambda}^*\)), Standard Errors (SE), Standardized Pattern Coefficients (\(\hat{\lambda}^*\)), and Percentage of Variance Accounted for (\(R^2\))

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1 = ME</th>
<th>Factor 2 = GSE</th>
<th>Factor 3 = TE</th>
<th>Factor 4 = CBE</th>
<th>Factor 5 = PCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\hat{\lambda}_{p,1})</td>
<td>(\hat{\lambda}_{p,1}^*)</td>
<td>(\hat{\lambda}_{p,2})</td>
<td>(\hat{\lambda}_{p,2}^*)</td>
<td>(\hat{\lambda}_{p,3})</td>
</tr>
<tr>
<td>me1</td>
<td>0.98</td>
<td>0.57</td>
<td>0.40</td>
<td>0.23</td>
<td>-0.04</td>
</tr>
<tr>
<td>me2</td>
<td>1.14</td>
<td>0.65</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>me3</td>
<td>1.05</td>
<td>0.61</td>
<td>0.02</td>
<td>0.01</td>
<td>0.40</td>
</tr>
<tr>
<td>me4</td>
<td>0.72</td>
<td>0.40</td>
<td>0.17</td>
<td>0.10</td>
<td>0.68</td>
</tr>
<tr>
<td>gse1</td>
<td>0.04</td>
<td>0.02</td>
<td>1.54</td>
<td>0.86</td>
<td>-0.07</td>
</tr>
<tr>
<td>gse2</td>
<td>0.45</td>
<td>0.27</td>
<td>0.82</td>
<td>0.49</td>
<td>0.14</td>
</tr>
<tr>
<td>gse3</td>
<td>-0.24</td>
<td>-0.12</td>
<td>1.72</td>
<td>0.87</td>
<td>0.12</td>
</tr>
<tr>
<td>gse4</td>
<td>0.17</td>
<td>0.09</td>
<td>0.99</td>
<td>0.53</td>
<td>0.45</td>
</tr>
<tr>
<td>te1</td>
<td>0.09</td>
<td>0.04</td>
<td>0.32</td>
<td>0.15</td>
<td>1.64</td>
</tr>
<tr>
<td>te2</td>
<td>0.08</td>
<td>0.04</td>
<td>0.13</td>
<td>0.06</td>
<td>1.76</td>
</tr>
<tr>
<td>te3</td>
<td>0.30</td>
<td>0.15</td>
<td>0.00</td>
<td>0.00</td>
<td>1.70</td>
</tr>
<tr>
<td>te4</td>
<td>-0.26</td>
<td>-0.16</td>
<td>-0.09</td>
<td>-0.06</td>
<td>1.22</td>
</tr>
<tr>
<td>cbe1</td>
<td>0.03</td>
<td>0.07</td>
<td>0.21</td>
<td>0.11</td>
<td>-0.09</td>
</tr>
<tr>
<td>cbe2</td>
<td>0.25</td>
<td>0.15</td>
<td>0.33</td>
<td>0.20</td>
<td>0.05</td>
</tr>
<tr>
<td>cbe3</td>
<td>1.06</td>
<td>0.51</td>
<td>-0.15</td>
<td>-0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>cbe4</td>
<td>0.75</td>
<td>0.49</td>
<td>0.07</td>
<td>0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>pce1</td>
<td>0.44</td>
<td>0.21</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>pce2</td>
<td>-0.05</td>
<td>-0.03</td>
<td>0.12</td>
<td>0.08</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Note. Statistically significant coefficients are in boldface. ME = motivation efficacy; GSE = game strategy efficacy; TE = technique efficacy; CBE = character building efficacy; PCE = physical conditioning efficacy.
were nonzero) and likely resulted in upwardly biased interfactor correlations in Model 8. A reasonable answer to Research Question 2 was as follows: a more parsimonious (restrictive) CFA model, informed by the a priori measurement theory (see Figure 1), did not offer a viable alternative to a more complex (flexible) ESEM model (see Figure 2). From this point forward Model 7 was the accepted model.

Research Question 3

Parameters of interest were the nonzero elements within the $\Lambda$ depicted in Figure 1. Parameter estimates from $\hat{\Lambda}$ in Model 7 were treated as the population values. This approach was considered a reasonable balance between a priori theory (i.e., $\Lambda$ in Figure 2 included the theoretically driven nonzero elements within the $\Lambda$ depicted in Figure 1) and the results from the previous research questions (e.g., Model 7 fits better than Model 8).

In the vast majority of cases (i.e., 16 of 18 $\lambda^*$), a relatively small $N$ (i.e., $\sim300$) provided sufficient power. In a few cases (i.e., $\lambda^*_{\text{cheat},CBE}$ and $\lambda^*_{\text{pce2},PCE}$), however, a larger $N$ (i.e., $\sim600$) was needed to provide sufficient power. These findings were consistent with the smaller estimates provided in Table 3 (i.e.,

<table>
<thead>
<tr>
<th>Table 4  Interfactor Correlation Matrix  $\Psi$ and Reliability Estimates from Exploratory Structural Equation Model (ESEM) with Geomin Rotation and Confirmatory Factor Analysis (CFA)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interfactor Correlation Matrix—ESEM (Model 5)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>ME</td>
</tr>
<tr>
<td>GSE</td>
</tr>
<tr>
<td>TE</td>
</tr>
<tr>
<td>CBE</td>
</tr>
<tr>
<td>PCE</td>
</tr>
</tbody>
</table>

<p>| <strong>Interfactor Correlation Matrix—CFA (Model 7)</strong> |</p>
<table>
<thead>
<tr>
<th></th>
<th>ME</th>
<th>GSE</th>
<th>TE</th>
<th>CBE</th>
<th>PCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>(.88)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSE</td>
<td>.81</td>
<td>(.90)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TE</td>
<td>.80</td>
<td>.82</td>
<td>(.92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBE</td>
<td>.88</td>
<td>.62</td>
<td>.59</td>
<td>(.88)</td>
<td></td>
</tr>
<tr>
<td>PCE</td>
<td>.81</td>
<td>.78</td>
<td>.81</td>
<td>.63</td>
<td>(.75)</td>
</tr>
</tbody>
</table>

*Note. ME = motivation efficacy; GSE = game strategy efficacy; TE = technique efficacy; CBE = character building efficacy; PCE = physical conditioning efficacy. Coefficient $H$ estimates are enclosed in parentheses.
Table 5 Confirmatory Factor Analytic Pattern Coefficients ($\hat{\lambda}$), Standard Errors (SE), Standardized Pattern Coefficients ($\hat{\lambda}^0$), and Percentage of Variance Accounted For ($R^2$)

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1 = ME</th>
<th>Factor 2 = GSE</th>
<th>Factor 3 = TE</th>
<th>Factor 4 = CBE</th>
<th>Factor 5 = PCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\lambda}_{p,1}$</td>
<td>SE</td>
<td>$\hat{\lambda}_{p,1}^0$</td>
<td>$\hat{\lambda}_{p,2}$</td>
<td>SE</td>
</tr>
<tr>
<td>me1</td>
<td>1.23</td>
<td>.09</td>
<td>0.78</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>me2</td>
<td>1.11</td>
<td>.08</td>
<td>0.74</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>me3</td>
<td>1.26</td>
<td>.09</td>
<td>0.78</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>me4</td>
<td>1.72</td>
<td>.16</td>
<td>0.86</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>gse1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.15</td>
<td>.08</td>
</tr>
<tr>
<td>gse2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.54</td>
<td>.12</td>
</tr>
<tr>
<td>gse3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.27</td>
<td>.09</td>
</tr>
<tr>
<td>gse4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.92</td>
<td>.16</td>
</tr>
<tr>
<td>te1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>te2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>te3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>te4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>cbe1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>cbe2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>cbe3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>cbe4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>pce1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>pce2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. Statistically significant coefficients are in boldface. ME = motivation efficacy; GSE = game strategy efficacy; TE = technique efficacy; CBE = character building efficacy; PCE = physical conditioning efficacy.
Coaching Efficacy and ESEM

A reasonable answer to Research Question 3 was as follows: an $N$ of 300 for the vast majority of parameters of interest and an $N$ of 600 for all parameters of interest.

**Research Question 4**

Model 9a (males: $\chi^2_R (72) = 113$, $p = .001$, RMSEA = .035, CFI = .996, and TLI = .992) and Model 9b (females: $\chi^2_R (72) = 78$, $p = .307$, RMSEA = .032, CFI = .997, and TLI = .993) both exhibited at least close fit to the data. The statistical nonsignificance of the exact fit test for females was likely attributable to the small sample of females. Model 10 (identification constraints only) exhibited close fit to the data: $\chi^2_R (157) = 189$, $p = .041$, RMSEA = .027, CFI = .997, and TLI = .995.

Model 11 (invariant pattern coefficients) exhibited marginally statistically significant worse fit than Model 10: $\Delta \chi^2_R (63) = 83$, $p = .049$. This finding presented an interesting dilemma. Unlike CFA (where some of the elements with $\Lambda$ can be invariant while other elements can be noninvariant), in ESEM all of the elements with $\Lambda$ have to be specified as either invariant or noninvariant. It was decided to retain Model 11 given the marginal statistical gain by allowing noninvariance for all pattern coefficients, a preference for model parsimony, and the close fit of Model 11: $\chi^2_R (220) = 263$, $p = .026$, RMSEA = .026, CFI = .997, and TLI = .995. This decision was consistent with rough guidelines for a nested model comparison which generally failed to provide evidence in favor of Model 10 as compared with Model 11: $\Delta$CFI = .000, $\Delta$TLI = .000, $\Delta$RMSEA = −.001. Model 12 (invariant pattern coefficients and thresholds) fit the data as well as Model 11: $\Delta \chi^2_R (18) = 17$, $p = .503$, $\Delta$CFI = .000, $\Delta$TLI = .001, $\Delta$RMSEA = −.002.

Model 13 (invariant pattern coefficients, thresholds, and residual covariance matrix) exhibited statistically significant worse fit than Model 12: $\Delta \chi^2_R (18) = 37$, $p = .005$; though the rough guidelines for a nested model comparison suggested that magnitude of this difference may not have been large: $\Delta$CFI = −.004, $\Delta$TLI = −.004, $\Delta$RMSEA = .010. The residual variance of the latent response variate underlying responses to gse1 (i.e., gse1*) and the residual variance of the latent response variate underlying responses to cbe4 (i.e., cbe4*) were identified as problematic constraints (see Table 1 for item text). In both cases the residual variance was larger for females (elaborated upon in the Discussion). Model 13b allowed these two residual variances to be noninvariant, while constraining all other residual variances to invariance, and this model fit the data as well as Model 12: $\Delta \chi^2_R (16) = 25$, $p = .067$, $\Delta$CFI = .000, $\Delta$TLI = .000, $\Delta$RMSEA = .001. Model 13b (factorial invariance except for two residual variances) exhibited close fit to the data: $\chi^2_R (254) = 298$, $p = .031$, RMSEA = .025, CFI = .997, and TLI = .996. A reasonable answer to Research Question 4 was that there was evidence for partial factorial invariance by coach’s gender.

**Post Hoc Research Questions**

**Research Question 5.** Model 14 (invariant latent means) exhibited statistically significant worse fit than Model 13b: $\Delta \chi^2_R (5) = 30$, $p < .001$, $\Delta$CFI = −.008, $\Delta$TLI = −.008, $\Delta$RMSEA = .018. Model 13b was accepted. The latent means that differed were GSE, 0.48, $z = 3.35$, $p = .001$, and TE, 0.55, $z = 3.55$, $p < .001$; both of the means were larger for females as compared with males.
Research Question 6. Model 15 (invariant factor covariance matrix) exhibited statistically significant worse fit than Model 13b, $\Delta \chi^2 (15) = 29, p = .017$, although the rough guidelines for a nested model comparison suggested that magnitude of this difference was not large: $\Delta \text{CFI} = -.001$, $\Delta \text{TLI} = .000$, $\Delta \text{RMSEA} = .000$. A reasonable answer to Research Question 6 was that there was evidence that the factor covariance matrix did not meaningfully differ by coach’s gender as per the rough guidelines for a nested model comparison.

Discussion

The purpose of the current article was to provide a substantive-methodological synergy of potential importance to future research in sport and exercise psychology. The substantive focus was to improve the measurement of coaching efficacy by developing a revised version of the CES (based, in part, on advances put forth in the CES II-HST) for head coaches of youth sport teams: the CES II-YST. The methodological focus was ESEM, a methodology that integrates the advantages of EFA and CFA within the general SEM. The synergy was a demonstration of how ESEM (as compared with CFA) may be used, guided by content knowledge, to develop (or confirm) a measurement model for the CES II-YST.

While the substantive focus and the methodological focus of the current article were both primarily concerned with measurement, ESEM was briefly presented in the Introduction as a broader methodological framework (see Equation 1 and Equation 2) within which the next generation of sport and exercise psychology research may evaluate fuller conceptual models. Almost any latent variable model that can be imposed within the SEM framework can also be imposed within ESEM framework (Asparouhov & Muthén, 2009). An advantage of the ESEM framework, in instances of insufficient a priori measurement model(s), is reduced likelihood of biased estimates wrought by misspecification of the measurement model(s). Studies in sport and exercise psychology that are not focused on measurement can still be adversely affected by misspecification of the measurement model(s) because misspecification of the measurement model(s) can produce biased path coefficients (Asparouhov & Muthén, 2009; Kaplan, 1988). The authors of this article, therefore, recommend that the ESEM framework should be strongly considered (as opposed to imposing only CFA measurement models) in subsequent studies when a priori measurement theory is insufficient to warrant a confirmatory approach.

What constitutes insufficient a priori measurement theory, and therefore should lead to a preference for a more flexible ESEM versus a more restrictive CFA (or vice versa), is an emerging area of research. A relevant finding from a simulation study by Myers, Ahn, and Jin (2011) was that in instances of only approximate model-data fit (defined as: $0.05 < \text{RMSEA} \leq 0.08$), ESEM may be expected to fit better than CFA and produce key parameter estimates that, on average, are at least as accurate as the parallel estimates from a CFA model. That said, as observed in the current study, imposing a more flexible ESEM model will generally not be without some costs. For example, from a conceptual perspective it is difficult to argue that the pattern coefficient matrix provided in Table 2 (ESEM) is as interpretable as is the pattern coefficient matrix provided in Table 5 (CFA). The authors of this article believe that it will generally be the case that as theoretical models more closely approximate true models, accepted theoretical models will frequently be more...
complex than initially anticipated (e.g., compare Figure 1 to Table 2). Another ESEM-related cost observed in the current study was the dichotomous choice with regard to (non)-invariance of the entire pattern coefficient matrix by coach’s gender; a wider range of choices would have been available under CFA. Simply, the flexibility of the ESEM model is paid for by the researcher giving up some control (e.g., me1 indicates only ME; rotational determinacy). The degree to which the costs of an ESEM outweigh the benefits wrought by the flexibility of an ESEM will need to be made on a case-by-case basis and will sometimes be reasonably debatable. It is our belief that CFA is generally a more desirable framework (given a sufficient degree of a priori measurement theory) than ESEM from a theoretical perspective. Simply, it is better to tell a statistical program what the true theoretical model is and then receive confirmatory feedback from the program, then it is to tacitly ask a inanimate statistical program to codevelop a theoretical model. In some cases, however, the less desirable ESEM framework will be a wiser choice than the CFA framework (e.g., the current study). Interestingly, in some cases even the ESEM framework may be viewed as too restrictive and automated causal discovery techniques may be more appropriate (e.g., Kalisch & Bühlmann, 2010; Landsheer, 2010; Rozeboom, 2008).

Development of CES II-YST was congruent with relevant research, was guided by content experts, and represented substantial revision to the CES via revision to the CES II-HST. Validity evidence was provided for the measurement model depicted in Figure 2 versus the measurement model depicted in Figure 1. Further, validity evidence was provided for close fit of the measurement model depicted in Figure 2 and for the veracity of imposing this model with a relatively small sample size ($N \sim 300$ for the vast majority of parameters of interest). Validity evidence was also provided for partial factorial invariance (Byrne, Shavelson, & Muthén, 1989) by coach’s gender. Empirical implications of partial factorial invariance (versus full factorial invariance), particularly for ordinal data, are largely unclear (Millsap & Kwok, 2004). The level of noninvariance observed in this study (161 of 163 measurement parameters were specified as invariant) was assumed to exert little practical impact on the ability to compare measures across coach’s gender. The degree to which this assumption was reasonable was a potential limitation of the current study. Another limitation of the current study was the small sample size for females ($n = 76$). The underrepresentation of females in the current study, while unfortunate, paralleled established trends in coaching (Acosta & Carpenter, 2006).

Two measurement parameters (residual variance of gse1* and the residual variance of cbe4*) were noninvariant by coach’s gender. In both cases the residual variance was larger for females, which implies that both items were less reliable indicators (i.e., as evidenced by variance accounted for) for females than males. It is interesting to note that Myers, Chase, Beauchamp, and Jackson (2010) provided similar evidence for gse1, make effective strategic decisions in pressure situations during competition, by athlete gender. Unlike gse1, cbe4, teach life lessons to your athletes through sport, was a new item. Future research (perhaps a targeted qualitative approach) that explores why these two items (particularly gse1) appear to be showing some signs of noninvariance by gender would be useful—particularly if each item can be revised in a relatively minor way that encourages invariance by gender. It is also important to recall that there was marginal statistical evidence for noninvariance of the pattern coefficient matrix. As the measurement model for the...
CES II-YST becomes more refined across time, and hence moves from the ESEM framework to the CFA framework, it will be important to determine if particular pattern coefficients are noninvariant by coach’s gender.

Post hoc research questions revealed two substantive findings in the nonmeasurement part of the model. First, there was evidence that the factor covariance matrix did not meaningfully differ by coach’s gender. This suggests that the degree to which latent coaching efficacy dimensions vary and covary with each other may be approximately equal for male and female coaches of youth sports. From a conceptual perspective, this finding suggests that any problems with discriminant validity based on the correlations among the coaching efficacy dimensions (ranged from .32 to .71) should be relatively constant for both male and female coaches. Perhaps more importantly, any discriminant validity problems associated with the CES II-YST (under the proposed ESEM model), appears to be equal to or less than that observed with the CES, which parallels findings from the CES II-HST versus the CES (Myers, Feltz, Chase, et al., 2008). From a methodological perspective, this finding encourages model parsimony in that 15 fewer parameters (5 variances and 10 covariances) may be estimated in future research.

The second substantive finding was with regard to latent mean differences on the dimensions of coaching efficacy by coach’s gender. Male and female coaches had approximately equal latent means with regard to motivation, character building, and physical conditioning efficacy. This set of findings was consistent with some previous research at other levels of competition: high school (e.g., Feltz et al., 1999; Malete & Feltz, 2000; Myers, Feltz, et al., 2011) and collegiate (e.g., Myers, Vargas-Tonsing, et al., 2005); and inconsistent with some other previous research: collegiate (Marback, Short, Short, & Sullivan, 2005) and mixed levels (Campbell & Sullivan, 2006). Female coaches had larger latent means on game strategy and technique efficacy as compared with male coaches. This set of findings was inconsistent with each of the studies previously cited in this paragraph (which found either nonsignificant differences or greater means for males as compared with females) and with Lee, Malete, and Feltz (2002) who studied a sample of youth coaches from Singapore. Comparing the results from the current study to previous studies is limited for at least two reasons (in addition to level coached and country). First, the current study compared latent means (i.e., accounting for measurement error) while all of the other studies cited (except for Myers, Feltz, et al.) used manifest means. Second, the current study used the CES II-YST while all of the other studies (except for Myers, Feltz, et al.) used the CES. With these limits in mind, however, we offer a speculative interpretation. We believe that it may be the case that a very talented group of adult females (with a lot of athletic experience relevant to technique and game strategy efficacy) may be willing and invited to coach at the youth sport level while being unwilling and/or less likely to be invited (as compared with males) to coach at higher levels of competition.

There is much work yet to be done to investigate the utility of coaching efficacy at the youth sport level. Close model-data fit for the CES II-YST implies little about the ability of resultant measures to relate to theoretically relevant external variables. While Feltz et al. (2009) provided evidence for proposed relationships between coach efficacy and a host of external variables (e.g., playing/coaching experience, athlete improvement, and social support) at the youth sport level, many other key relations await investigation (e.g., outcomes of coaching efficacy). A related line
of research that also may be important is to develop an analog to the CES II-YST that allows youth athletes to evaluate their head coach’s coaching competency, and, to link these measures to external variables within broader models of coaching effectiveness.

There are an estimated 20–35 million American children playing competitive youth sports and this number has risen significantly over the last 10–20 years (Ewing & Seefeldt, 2002). Given the vast number of youth sport participants and the impact that coaching behavior can have on young athletes, improving the measurement of key coaching effectiveness constructs at the youth sport level is a vital area of research. Development of the CES II-YST may facilitate much needed research with youth sport coaches that examine coaching efficacy and potential outcomes for coaches, athletes, and teams. Unique to the CES II-YST are items that measure efficacy dimensions designed specifically for coaching children ages 8–13 years—an age group for which improved measurement of coaching efficacy has been advocated (Myers, Feltz, Chase, et al., 2008).

Notes

1. This a more restrictive approach to simple structure than advocated by Thurstone (1947) as detailed in Browne (2001). A perfect simple structure has also been referred to as independent clusters confirmatory factor analysis (Marsh et al., 2009).

2. For the single-group case, EFA (without means) only within the ESEM framework is analogous to traditional EFA (Asparouhov & Muthén, 2009).

3. This recommendation was based on simulation research with SEM models (e.g., CFA) before the advent of ESEM. The few published applications of ESEM that the authors are aware of (e.g., Marsh et al., 2009; Marsh et al., 2010) have generally treated ordinal data as conditionally normal and continuous with a correction for non-normality. The degree to which CVM under WLSMV estimation for ordinal data represents a methodological advantage (e.g., correcting for attenuation in parameter estimates that may result due to the coarseness of the data) as compared with a robust normal theory estimator with respect to ESEM is unknown.

4. The DIFFTEST command was used to obtain an accurate test. Mathematical details of this approach (e.g., scaling and degrees of freedom) are available in Asparouhov and Muthén (2010).

5. Marsh et al. (2009) put forth a taxonomy of 13 multiple group invariance tests. Models in the current study related to this taxonomy are as follows: Model 10 = configural invariance; Model 11 = weak factorial invariance; Model 12 = strong factorial invariance; Model 13 = strict factorial invariance; Model 14 = manifest mean invariance; Model 15 = complete factorial invariance.

6. The use of manifest means brings up an important practical consideration: how to derive measures from responses to the CES II-YST. Results from this study indicate that summing responses from items intended to measure a particular dimension would likely be suboptimal. An optimal approach would be to impose an ESEM model with or without additional covariates.

References


*Manuscript submitted*: April 18, 2011

*Revision accepted*: July 11, 2011
Appendix

The first equation is the measurement model (Bollen, 1989):

\[ y_{p \times 1} = \nu_{p \times 1} + \Lambda_{p \times m} \eta_{m \times 1} + K_{p \times q} x_{q \times 1} + \varepsilon_{p \times 1} \]  

(1)

where

- \( p \) is the number of continuous observed dependent variables
- \( m \) is the number of continuous latent variables
- \( q \) is the number of observed independent variables
- \( y \) is a vector of continuous observed dependent variables
- \( \nu \) is a vector of intercepts
- \( \Lambda \) is a matrix of pattern coefficients
- \( \eta \) is a vector of continuous latent variables
- \( K \) is a matrix of regression coefficients: \( Y \) on \( X \)
- \( x \) is a vector of observed independent variables
- \( \varepsilon \) is a vector of residuals for \( Y \)
- \( \Theta \) is a \( p \times p \) covariance matrix for \( \varepsilon \)

The second equation is the latent variable model (Bollen):

\[ \eta_{m \times 1} = \alpha_{m \times 1} + \beta_{m \times m} \eta_{m \times 1} + \Gamma_{m \times q} x_{q \times 1} + \zeta_{m \times 1} \]  

(2)

where

- \( \alpha \) is a vector of intercepts
- \( \beta \) is a matrix of regression coefficients: \( \eta \) on \( \eta \)
- \( \Gamma \) is a matrix of regression coefficients: \( \eta \) on \( X \)
- \( \zeta \) is a vector of residuals for \( \eta \)
- \( \Psi \) is a \( m \times m \) covariance matrix for \( \zeta \)