*Journal of Sport & Exercise Psychology,* 2012, 34, 737-765
© 2012 Human Kinetics, Inc.

JOURNAL OF
SPORT & EXERCISE
PSYCHOLOGY
Official Journal of NASPSPA
www.JSEP-Journal.com
ORIGINAL RESEARCH

# Development of, and Initial Validity Evidence for, the Referee Self-Efficacy Scale: A Multistudy Report

**Nicholas D. Myers,[1] Deborah L. Feltz,[2] Félix Guillén,[3] and Lori Dithurbide[4]**
[1]University of Miami; [2]Michigan State University; [3]University of Las Palmas de Gran Canaria; [4]Saint Mary's University, Halifax

The purpose of this multistudy report was to develop, and then to provide initial validity evidence for measures derived from, the Referee Self-Efficacy Scale. Data were collected from referees ($N = 1609$) in the United States ($n = 978$) and Spain ($n = 631$). In Study 1 ($n = 512$), a single-group exploratory structural equation model provided evidence for four factors: game knowledge, decision making, pressure, and communication. In Study 2 ($n = 1153$), multiple-group confirmatory factor analytic models provided evidence for partial factorial invariance by country, level of competition, team gender, and sport refereed. In Study 3 ($n = 456$), potential sources of referee self-efficacy information combined to account for a moderate or large amount of variance in each dimension of referee self-efficacy with years of referee experience, highest level refereed, physical/mental preparation, and environmental comfort, each exerting at least two statistically significant direct effects.

***Keywords:*** sports officials, sources of sport confidence, sport refereed, exploratory structural equation modeling, target rotation

The purpose of this multistudy report was to develop, and then to provide initial validity evidence for measures derived from, the Referee Self-Efficacy Scale (REFS). The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) included at least four relevant guidelines for this investigation. First, a conceptual framework was provided. Second, the development process for the REFS was described. Third, competing a priori measurement theories were put forth to explain responses to

Nicholas D. Myers is with the Department of Educational and Psychological Studies, University of Miami, Coral Gables, FL. Deborah L. Feltz is with the Department of Kinesiology, Michigan State University, East Lansing. MI. Félix Guillén is with the Department of Psychology and Sociology, University of Las Palmas de Gran Canaria, Spain. Lori Dithurbide is now with the School of Health and Human Performance, Dalhousie University, Halifax, Nova Scotia, Canada.

the REFS (i.e., internal validity). Fourth, proposed relationships between measures derived from the REFS and other theoretically relevant variables were tested (i.e., external validity).

## A Conceptual Framework

Referee self-efficacy was conceptualized within self-efficacy theory (Bandura, 1997), and more specifically, self-efficacy in sport (Feltz, Short, & Sullivan, 2008). Self-efficacy judgments are domain-specific beliefs held by individuals about their ability to successfully execute differing levels of performance given certain situational demands. Sports officials must execute multiple tasks, under pressure, to perform their roles successfully in a competition and not make errors in judgment. For instance, sports officials must evaluate and judge actions that take place during the match, make fast decisions, manage the game, pay attention to multiple aspects of the game, keep order, and solve disputes, all under socially evaluative conditions (Tuero et al., 2002). Lack of efficacy can lead to lapses in attention, errors in judgment, delayed reactions, and eventual stress and burnout (Guillén & Feltz, 2011).

Within sport psychology, there is ample evidence that given sufficient incentive to perform and requisite skills for a given task, efficacy beliefs generally are important for athletes (e.g., Jackson, Beauchamp, & Knapp, 2007; Moritz, Feltz, Fahrbach, & Mack, 2000), teams (e.g., Feltz & Lirgg, 1998; Spink, 1990a), and coaches (e.g., Feltz, Chase, Moritz, & Sullivan, 1999) as reviewed by Feltz et al. (2008). Within each of these specific populations within sport, advances in the relevant research have been guided by development of population-specific conceptual (e.g., Feltz, 1982; Feltz et al., 1999; Lent & Lopez, 2002; Spink, 1990b) and measurement models (e.g., Feltz et al., 1999; Short, Sullivan, & Feltz, 2005). Guillén and Feltz (2011) argued that referees are an important population in sport that has largely been ignored in terms of their efficacy beliefs for referee performance. Therefore, conceptual and measurement models are needed for guiding research in this area.

A preliminary conceptual model of referee self-efficacy was put forth by Guillén and Feltz (2011).[1] Referee self-efficacy was defined as the extent to which referees believe they have the capacity to perform successfully in their job. Borrowing from self-efficacy theory and self-efficacy research in sport, Guillén and Feltz proposed that highly efficacious referees should be more accurate in their decisions, more effective in their performance, more committed to their profession, have more respect from coaches, administrators, and other officials and suffer less stress from officiating than less efficacious referees. Potential outcomes of referee self-efficacy were not a focus of the current multistudy report.

Proposed sources of referee self-efficacy in Guillén and Feltz (2011) were consistent with Bandura's (1997) sources of efficacy information categories and included subscales from the Sources of Sport Confidence Questionnaire (SSCQ; Vealey, Hayashi, Garner-Holman, & Giacobbi, 1998) because the SSCQ itself can be conceptualized within self-efficacy theory. Proposed sources of referee self-efficacy included mastery experiences (analogous to Bandura's past performance accomplishments category), significant others (analogous to Bandura's verbal persuasion category and included the social support subscale of the SSCQ), physical and mental preparation (included the physical/mental preparation subscale of the SSCQ and mental aspects are analogous to Bandura's emotional arousal category), and partner qualifications (included the environmental comfort and situational favorableness subscales of the SSCQ). Mastery experiences (e.g., years of referee experience, highest level refereed) were hypothesized to be the strongest source of referee self-efficacy information (Guillén & Feltz). One potential source of efficacy information category not mentioned in Guillén and Feltz, but consistent with self-efficacy theory, is vicarious experience, which has a subscale in the SSCQ. The proposed sources of referee self-efficacy have yet to be tested.

Proposed dimensions of referee self-efficacy outlined in Guillén and Feltz (2011) included game knowledge, strategic skills, decision-making skills, psychological skills, communication/control of game, and physical fitness. Guillén and Feltz cautioned that their conceptual model should serve as only a starting point for subsequent research to (dis)confirm and modify. For instance, specific operational definitions for each of the proposed dimensions of referee self-efficacy were not put forth by Guillén and Feltz. Items were not developed by Guillén and Feltz to indicate the dimensions. Further development of an explicit measurement model for referee self-efficacy was an objective of the current study and informed specific research questions in the current study. Development of a measurement model for referee self-efficacy will be described in the next section. Specific research questions for the current study were informed, in part, by the development of a measurement model for referee self-efficacy and, therefore, will be provided at the end of the this introductory section.

## Development of the REFS

Development of the REFS was accomplished in the current study via an iterative process guided by four experts in the psychosocial aspects of sport and physical activity. Within this expert group, two of the members also had expertise in the measurement of self-efficacy in sport. The expert group critically reviewed the relevant conceptual (e.g., Feltz et al., 2008) and measurement literature (e.g., Feltz & Chase, 1998) and considered themes from a focus group with nine referees of soccer from the Midwestern region of the United States of America (US).[2] An implicit assumption of the focus group was that information gleaned from referees of soccer may generalize to referees of other team sports. The composition and results of the focus group both were detailed by Guillén and Feltz (2011).

Referee self-efficacy was defined as the extent to which a referee believes that he or she has the ability to successfully officiate a competition. The operational definition proposed in the current study was slightly different than what was provided in Guillén and Feltz (2011, p. 1) who defined referee efficacy as “the extent to which referees believe they have the capacity to perform successfully in their job.” The changes in the current study were considered minor (e.g., further defining the task for which efficacy beliefs were sought); were viewed as increasing the precision of the operational definition; and more closely matched the operational definition of other self-efficacy constructs in sport (e.g., Myers, Chase, Pierce, & Martin, 2011).

Four first-order dimensions of referee self-efficacy were conceptualized. Game knowledge (GK) was defined as the confidence that a referee has in his/her knowledge of his/her sport. Decision making (DM) was defined as the confidence that a referee has in his/her ability to make decisions. Pressure (PR) was defined as the confidence that a referee has in his/her ability to be uninfluenced by pressure.

Communication (CM) was defined as the confidence that a referee has in his/her ability to communicate effectively. Two of the dimensions sketched in Guillén and Feltz (2011), psychological skills and control of the game, were collapsed into a single dimension in the current study: pressure. This change was viewed as a better way to consolidate the feedback from the focus group for the purpose of instrument development.

The REFS was developed for referees of team sports. Level of competition refereed was not delimited. Consistent with some recent literature on the measurement of self-efficacy in sport (e.g., Myers, Chase et al., 2011), each dimension of the REFS was defined by only a few items. There were multiple justifications for this decision. Practically, recruitment for data collection can be a difficult task and one that is 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 referee self-efficacy. Empirically, self-efficacy scales in sport have a stable history of at least a moderately high pattern coefficient value on the intended factor (e.g., Myers, Feltz, Chase, Reckase, & Hancock, 2008).

Item development and the proposed underlying factor structure of the REFS were based on a feedback loop between substantive theory and insight provided by the focus group. Table 1 provides the text for the REFS items. The item stem was, "in relation to the primary sport(s) that you referee, how confident are you in your ability to. . . ." A five-category rating scale was initially implemented based on previous research (Myers, Wolfe, & Feltz, 2005).

***A Priori Measurement Theories.*** The initial a priori measurement theory for the REFS was depicted from an ESEM (Asparouhov & Muthén, 2009) perspective in Figure 1 and from a CFA perspective in Figure 2. As is common in CFA, the model depicted in Figure 2 initially assumed a perfect simple structure (i.e., variable complexity, $vc = 1$ for each variable). When sufficient a priori measurement theory exists, CFA is preferred over ESEM due to its computational efficiency. As is common in ESEM, the model in Figure 1 allowed for a complex structure (i.e., $vc$ generally is free to equal the number of factors, $m$, 4 in this case). When sufficient a priori measurement theory does not exist, the more complex ESEM is preferred over CFA due to the reduced likelihood of biased estimates wrought by misspecification of the measurement model (Asparouhov & Muthén). In preliminary validity studies it is difficult to know if sufficient a priori measurement theory exists, and thus, using both perspectives can be a reasonable approach (see Myers, Chase, et al., 2011, for a thorough discussion). Five measurement residual covariances were hypothesized based on the wording of the items (see Figure 1). The rationale for each of these minor factors is provided as a specific note in Table 1.

***N and the REFS.*** After a preliminary measurement model was accepted an important aspect of the development of the REFS was determining what $N$ would be needed to achieve a particular level of power for subsequent studies. Rules of thumb (e.g., $N \geq 200$) for determining adequate $N$ for a particular application of factor analysis with real data are known to be of limited use (Marsh, Hau, Balla, & Grayson, 1998). Monte Carlo methods can be used in real data analysis to decide on $N$ and to estimate power (Muthén & Muthén, 2002). The importance of implementing such an approach in validity studies in exercise and sport, as opposed to relying on rules of thumb, has been demonstrated (Myers, Ahn, & Jin, 2011).

**Table 1 Operational Definitions and Items for the Referee Self-Efficacy Scale (REFS)**

Game knowledge (GK): confidence that a referee has in his/her knowledge of their sport

gk1: understand the basic strategy of the game

gk2: understand all the rules of your sport (with cm4)[a] (** with cm2)[a]

gk3: understand proper officiating mechanics

dm1: make critical decisions during competition

dm3: make quick decisions

cm2: communicate effectively with other referees

Decision making (DM): confidence that a referee has in his/her ability to make decisions during competition

dm1: make critical decisions during competition (with cm4)[a] (** with dm2)[b]

dm2: be firm in your decisions (** with cm4)[a]

dm3: make quick decisions (** with gk2)[b]

pr3: uninfluenced by pressure from coaches

cm2: communicate effectively with other referees

Pressure (PR): confidence that a referee has in his/her ability to be uninfluenced by pressure

pr1: uninfluenced by pressure from players (with cm3)[c] (** with dm1)[c]

pr2: uninfluenced by pressure from spectators (** with gk2 and dm3)[b]

pr3: uninfluenced by pressure from coaches (with cm1)[d] (** with dm2, gk1, and gk3)[d]

dm1: make critical decisions during competition

dm2: be firm in your decisions

Communication (CM): confidence that a referee has in his/her ability to communicate effectively

cm1: communicate effectively with coaches (with cm3)[e]

cm2: communicate effectively with other referees (** with gk1 and cm4)[a]

cm3: communicate effectively with players (** with gk1 and pr2)[c]

cm4: communicate effectively with auxiliary game personnel

gk1: understand the basic strategy of the game

*gk3: understand proper officiating mechanics

*dm1: make critical decisions during competition

*Secondary pattern coefficient specified post hoc.

**Residual covariance specified post hoc.

[a] The minor factor was conceptualized as interaction with other game personnel.

[b] The minor factor was conceptualized as game management.

[c] The minor factor was conceptualized as interaction with players.

[d] The minor factor was conceptualized as interaction with coaches.

[e] The minor factor was conceptualized as interaction with primary participants.

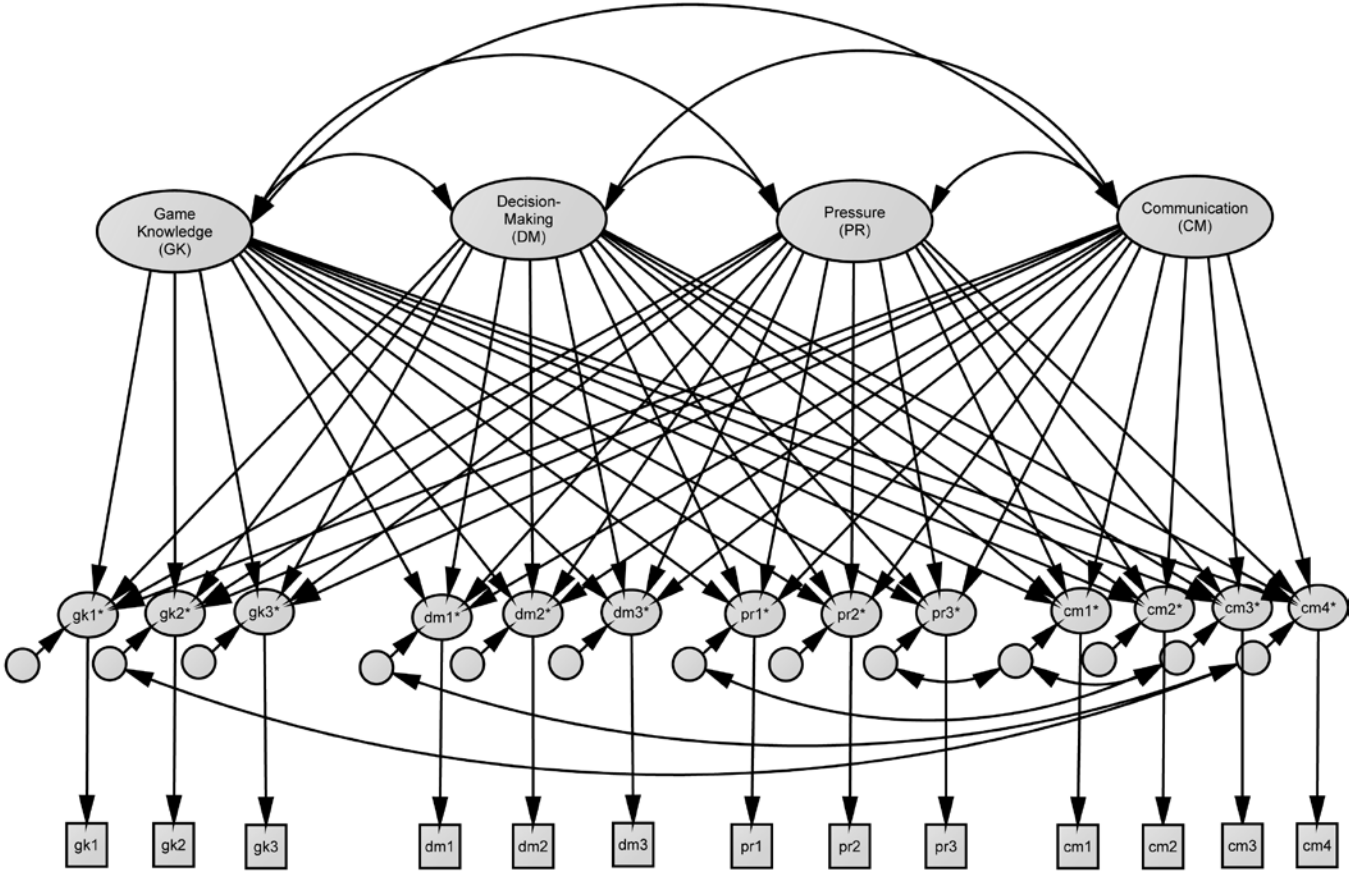


**Figure 1** — Initial a priori measurement theory for the REFS from an ESEM perspective.

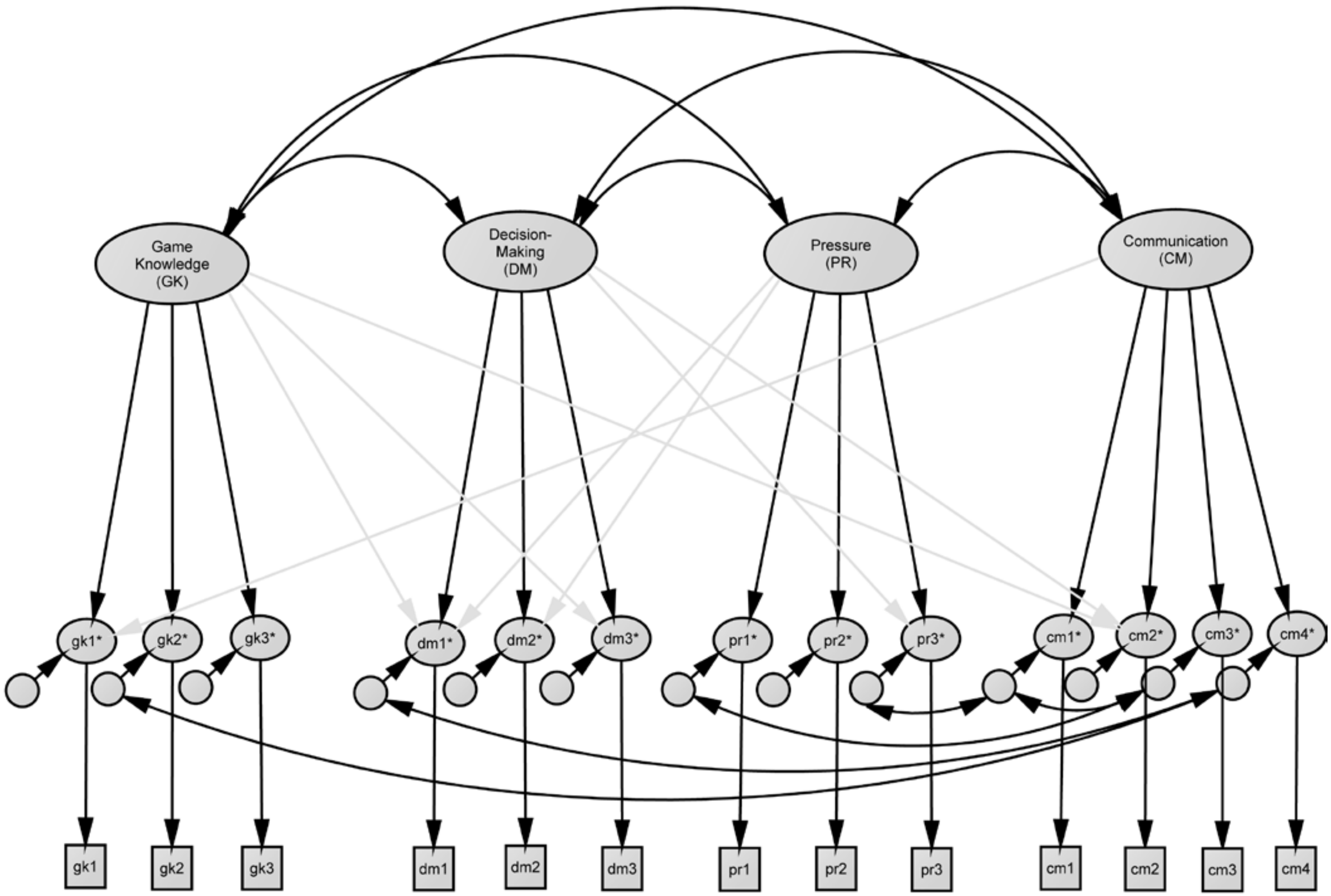


**Figure 2** — Initial a priori measurement theory for the REFS from a CFA perspective. The grayed arrows were not hypothesized before Study 1.

***Measurement Invariance.*** Measurement invariance is necessary to establish that measures are comparable across groups (Millsap, 2011). Providing evidence for measurement invariance across groups for which a new instrument is intended is an important facet of validity evidence (AERA, APA, NCME, 1999). The intended population for the REFS spanned several potentially important grouping variables (e.g., country, level of competition, team gender, and sport refereed). Within the self-efficacy in sport literature there is evidence for measurement non-invariance by country, level of competition, and gender (e.g., Myers, Chase, Beauchamp, & Jackson, 2010; Myers, Wolfe, Feltz, & Penfield, 2006). Evidence for measurement invariance by sport refereed, however, does not exist in the self-efficacy literature, which makes it unclear if it is reasonable to assume that REFS measures would be comparable across relevant subgroups.

## Research Questions

Five research questions were investigated within a multistudy design. The first four research questions investigated the internal validity of REFS measurement model(s). The fifth research question investigated the external validity of measures derived from the REFS.

***Research Question 1.*** How many factors were warranted to explain responses to the REFS? As depicted in Figures 1 and 2, the a priori hypothesis was four factors.

***Research Question 2.*** Could a more restrictive CFA that was primarily informed by a priori measurement theory offer a viable alternative to a more flexible ESEM?

***Research Question 3.*** What would be the minimum necessary *N* for a desired level of power with regard to the accepted measurement model for the REFS for subsequent studies?

***Research Question 4.*** Was there evidence for factorial invariance by country, level of competition, team gender, and sport refereed?

***Research Question 5.*** Could a set of theoretically defensible potential sources of efficacy information (i.e., years of referee experience, highest level refereed, and SSCQ dimensions) combine to account for a meaningful amount of variance in each dimension of referee self-efficacy? Based on self-efficacy theory, the a priori hypothesis was that the identified set of sources of efficacy information would account for a meaningful amount of variance in each dimension of referee self-efficacy. A priori hypotheses regarding the unique ability of particular sources to predict particular dimensions of referee self-efficacy were not specified due to a lack of research in this area.

# General Method

## Data Collection

***Procedure.*** An institutional review board provided necessary permission. Data were collected from referees of team sports in the US and Spain over 15 months. Data were collected both in-person and electronically to provide access to a large and diverse sample of referees. For the US data, a state high school athletic association and referee contacts assisted with the distribution of the link to the online survey through referee e-mail lists. For the Spain data, questionnaires were administered with the help of local referee associations. Informed consent was obtained from all participants. Referees were assured of confidentiality for their responses.

***Participants.*** Total sample size was 1609 ($n_{\text{US}}$ = 978, $n_{\text{Spain}}$ = 631) representing 15 different team sports. Sports that comprised at least 5% of the sample included basketball (*n* = 555), soccer (*n* = 415), volleyball (*n* = 215), roller hockey (*n* = 89), and American football (*n* = 77). Highest level of competition refereed included four groups: youth, recreation, and/or junior varsity (*n* = 283), high school or non-Division I university (*n* = 645), Division I university and/or semiprofessional (*n* = 346), and professional and/or international (*n* = 65). Referees who refereed male teams primarily (*n* = 510), female teams primarily (*n* = 243), and male and female teams primarily (*n* = 798) were represented. A majority of participants identified themselves as male (*n* = 1454). Age of referee ranged from 18 to 79 years (*M* = 38.38, *SD* = 13.98).

***REFS Descriptives.*** Responses were not observed in each category for two items—a requirement of some subsequent analyses. In both items (dm3, *make quick decisions*, and cm4, *communicate effectively with auxiliary game personnel*), the first category (no confidence) was the response option without observations. Across items and cases, only 0.34% of responses were observed in the first category. Responses in the first category were collapsed into the second category for each item resulting in a four-category rating scale structure. This post hoc collapsing decision, along with the resultant four-category rating scale (i.e., low, moderate, high, and complete confidence) was consistent with previous research (Myers, Feltz, & Wolfe, 2008). Within each rating scale category and across items, mean percentage of observed responses was 1.63% (*SD* = 0.92%) for the first category, 11.24% (*SD* = 4.65%) for the second category, and 39.30% (*SD* = 5.14%) for the third category, and 47.54% (*SD* = 9.32%) for the fourth category.

***Missing Data.*** Missing data comprised 23.9% of cells in the raw data matrix. Almost all (i.e., 94.6%) of the missing data were observed on the SSCQ. These SSCQ missing data were missing by design as they were purposely collected only from a subset of referees within Study 3. Missing data were handled from this point forward with the relevant default approach (e.g., pairwise present) in M*plus* 6 (Muthén & Muthén, 1998–2010) under the assumption of missing completely at random (Schafer & Graham, 2002).

## Statistical Modeling

***Ordinal Data.*** Figures 1 and 2 both imply categorical variable methodology (CVM; Muthén, 1984). In cases where the number of response options is less than five, Finney and DiStefano (2006) suggest using CVM with weighted least squares mean- and variance-adjusted (WLSMV) estimation. The REFS data were modeled as ordinal using CVM under WLSMV estimation.

***Model–Data Fit, Reliability, and Potential Sources of Misspecification.*** Indexes of model-data fit considered were $\chi^2_R$, RMSEA, CFI, and TLI. SRMR is unavailable under WLSMV estimation. Heuristic classifications model–data fit (exact, close, etc.) were consistent both with Hu and Bentler (1999) and with cautions against overreliance on empirical model–data indices at the expense of substantive considerations (Marsh, Hau, & Wen, 2004). Construct reliability was measured with coefficient *H* (Hancock & Mueller, 2001). At key intervals, post hoc investigations of modification indices (MIs) were conducted to determine possible locations of model misspecification (Saris, Satorra, & van der Veld, 2009).

# Study 1

The first three research questions were investigated in Study 1. Data ($N = 512$) were collected in the US ($n = 325$) and Spain ($n = 187$).

## Methods

***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., $m = 1$) as compared with the next more complex alternative ESEM (i.e., $m = 2$). Five sequential models were fit by systematically increasing $m$: Model 1 ($m = 1$) to Model 5 ($m = 5$). Nested models were compared with the change in the likelihood ratio $\chi^2$ (robust) test, $\Delta\chi^2_R$, with the DIFFTEST command (Asparouhov & Muthén, 2010). The approach taken in Step 2 is susceptible to overfactoring and inflated Type I error (Hayashi, Bentler, & Yuan, 2007), particularly when both models being compared are misspecified (Yuan & Bentler, 2004).[3] Therefore, alpha was set to .01 for these comparisons ($\alpha = .05$ otherwise) and the interpretability of the estimated rotated pattern matrix, $\hat{\Lambda}^*$, was also considered when deciding which model to accept. Oblique target rotation (Browne, 2001) was selected because it was designed to rotate the estimated pattern matrix, $\hat{\Lambda}$, toward a partially specified (i.e., targeted) matrix, which can take advantage of a priori content knowledge.

Given the weaknesses of the $\Delta\chi^2_R$ with real data and imperfect theories (Yuan & Bentler, 2004) and the utility of strict adherence to null hypothesis testing with regard to the assessment of model-data fit in general (Marsh et al., 2004), a set of guidelines were also used to judge the magnitude of change in model-data fit for nested models (e.g., $\text{CFI}_{\text{simple}} - \text{CFI}_{\text{complex}} = \Delta\text{CFI}$). Consistent with Marsh et al. (2010), $\Delta\text{CFI} \leq -.01$, $\Delta\text{TLI} \leq .00$, and $\Delta\text{RMSEA} \geq .015$, was interpreted as evidence in favor of the more complex model. From this point forward, ΔCFI, ΔTLI, and ΔRMSEA were collectively referred to as guidelines for a nested model comparison.

***Secondary Pattern Coefficients.*** Before addressing Research Question 2, $\hat{\Lambda}^*$ from the accepted ESEM solution was inspected for statistically significant and theoretically defensible secondary pattern coefficients (i.e., "cross-loadings"). A secondary pattern coefficient can be thought of as a nonzero "loading" on a factor that an item was not initially intended to measure in addition to a nonzero "loading" on the factor that the item was intended to measure. ESEM, not CFA, is typically the better framework for detecting secondary pattern coefficients (Browne, 2001). Forcing nonzero pattern coefficients to equal zero in a CFA model 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). Including secondary pattern coefficients in a post hoc manner, however, is susceptible to capitalization on chance and should be considered tentative before replication (MacCallum, Roznowski, & Necowitz, 1992).

***Research Question 2.*** The second research question was answered via $\Delta\chi^2_R$, $\Delta$CFI, $\Delta$TLI, and $\Delta$RMSEA because the parametrically simpler CFA in Figure 2 (i.e., Model 6) was nested within the more complex ESEM displayed in Figure 1 (i.e., Model 4). As displayed in Figure 1, ESEM imposed fewer restrictions on $\mathbf{\Lambda}$ than the CFA displayed in Figure 2.

***Research Question 3.*** After accepting a measurement model for responses to the REFS, a minimum necessary $N$ for a desired level of power was determined for a set of parameters of interest ($\mathbf{\Lambda}$, where $\lambda_i$ was a particular parameter of interest) for subsequent studies with an approach that was consistent with Myers, Ahn et al. (2011). Monte Carlo methods were used to determine the minimum $N$ at which each $H_0 : \lambda_i = 0$ was rejected in at least 80% of the replications ($\alpha = .05$). Number of replications was set to 10,000.

## Results

***Research Question 1.*** The null hypothesis for exact fit was rejected through four-factor extraction but failed to be rejected for the five-factor solution (see Table 2). The five-factor solution was judged not interpretable, as no element in the fifth column of $\hat{\mathbf{\Lambda}}^*$ was statistically significant (i.e., an empirical indication of over-factoring; Hayashi et al., 2007). Moreover, the four-factor solution (i.e., Model 4) fit the data as well as the five-factor solution (i.e., Model 5), $\Delta\chi^2_R(9) = 21, p = .013$. Model 4 exhibited close fit: $\chi^2_R(27) = 43, p = .029$, RMSEA = .033, CFI = .997, and TLI = .992, and was accepted. Interfactor correlations ranged from $\hat{\psi}^*_{\text{DM,PR}} = .40$ to $\hat{\psi}^*_{\text{DM,CM}} = .70$, whereas reliability ranged from $\hat{H}_{\text{DM}} = .86$ to $\hat{H}_{\text{PR}} = .91$. A reasonable answer to Research Question 1 was that four ESEM factors explained responses to the REFS.

***Secondary Pattern Coefficients.*** Elements within $\hat{\mathbf{\Lambda}}^*$ from Model 4 were generally consistent with a priori expectations (compare Table 3 to Figure 2). Note that 9 of the 39 elements within $\hat{\mathbf{\Lambda}}^*$ that were estimated in the ESEM approach, and inconsistent with the a priori portion of Figure 2 (e.g., $\hat{\lambda}^*_{\text{gk1,CM}}$), were statistic ally significant. Most (i.e., 8 of 9) of these secondary pattern coefficients were viewed as potentially theoretically defensible and were tentatively accepted (see grayed arrows in Figure 2 and Table 1 for item content).

***Research Question 2.*** The CFA depicted in Figure 2 (i.e., Model 6) fit the data as well as the accepted ESEM (i.e., Model 4), $\Delta\chi^2_R(19) = 16, p = .641$ (see Table 2). Model 6 exhibited exact fit: $\chi^2_R(46) = 52, p = .246$. Nonzero elements within $\hat{\mathbf{\Lambda}}$ from Model 6 were consistent with expectations (compare Table 4 to Figure 2). Interfactor correlations ranged from $\hat{\psi}^*_{\text{DM,PR}} = .29$ to $\hat{\psi}^*_{\text{DM,CM}} = .72$, whereas reliability ranged from $\hat{H}_{\text{DM}} = .84$ to $\hat{H}_{\text{PR}} = .91$. A reasonable answer to Research Question 2 was that the CFA model, based on a priori measurement theory plus eight secondary pattern coefficients, offered a viable alternative to the more complex ESEM.

**Table 2 Key Results from Study 1, Research Question 1 (How Many ESEM Factors?) and Research Question 2 (CFA Versus ESEM)**

| | Question 1: Number of Factors (*m*) | | | | | | Question 1: *m* – 1 Versus *m* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **$\chi^2$(*df*)** | **Par** | **RMSEA** | **$CI_{90\%}$** | **CFI** | **TLI** | **Complex** | **$\Delta\chi^2$($\Delta$*df*)** | **ΔRMSEA** | **ΔCFI** | **ΔTLI** |
| Model 1: *m* = 1 | 699(60)*** | 57 | .144 | [.135, .154] | .879 | .843 | Model 2 | 312(12)*** | .052 | –.082 | –.094 |
| Model 2: *m* = 2 | 254(48)*** | 69 | .092 | [.081, .103] | .961 | .937 | Model 3 | 109(11)*** | .028 | –.024 | –.032 |
| Model 3: *m* = 3 | 115(37)*** | 80 | .064 | [.051, .078] | .985 | .969 | Model 4 | 61(10)*** | .030 | –.012 | –.023 |
| Model 4: *m* = 4 | 43(27)* | 90 | .034 | [.011, .052] | .997 | .992 | Model 5 | 21(9)* | .013 | –.002 | –.005 |
| Model 5: *m* = 5 | 22(18) | 99 | .021 | [.000, .047] | .999 | .997 | | | | | |
| | Question 2: Confirmatory Factor Analysis (CFA) Versus Exploratory Structural Equation Modeling (ESEM) | | | | | | | | | | |
| **Model** | **$\chi^2$(*df*)** | **Par** | **RMSEA** | | **CFI** | **TLI** | **Complex** | **$\Delta\chi^2$($\Delta$*df*)** | **ΔRMSEA** | **ΔCFI** | **ΔTLI** |
| Model 6: CFA | 52(46) | 71 | .016 | [.000, .035] | .999 | .998 | Model 4 | 16(19) | –.018 | .002 | .001 |

*Note. m* = number of factors. Par = number of parameters estimated. Complex = more complex model that a simpler nested model was compared to. Δ = change.

**p* < .05. ***p* < .01. ****p* < .001.

**Table 3 Study 1, Model 4: Target-Rotated Pattern Coefficients ($\Lambda$), Standard Errors (*SE*), Standardized Pattern Coefficients ($\Lambda^0$), and Percentage of Variance Accounted for ($R^2$)**

| | Factor 1 = GK | | | Factor 2 = DM | | | Factor 3 = PR | | | Factor 4 = CM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\lambda^*_{p1}$ | *SE* | $\lambda^*_{p1}{}^0$ | $\lambda^*_{p2}$ | *SE* | $\lambda^*_{p2}{}^0$ | $\lambda^*_{p3}$ | *SE* | $\lambda^*_{p3}{}^0$ | $\lambda^*_{p4}$ | *SE* | $\lambda^*_{p4}{}^0$ | $R^2$ |
| gk1 | **0.83** | .14 | **0.56** | –0.09 | .15 | –0.06 | –0.12 | .09 | –0.08 | **0.55** | .13 | **0.37** | .55 |
| gk2 | **1.43** | .25 | **0.80** | –0.01 | .20 | –0.01 | 0.01 | .12 | 0.01 | 0.08 | .15 | 0.04 | .68 |
| gk3 | **1.11** | .11 | **0.75** | 0.00 | .01 | 0.00 | 0.00 | .01 | 0.00 | –0.02 | .05 | –0.01 | .55 |
| dm1 | **0.38** | .13 | **0.24** | **0.74** | .17 | **0.48** | **0.39** | .12 | **0.26** | –0.13 | .08 | –0.09 | .58 |
| dm2 | 0.00 | .02 | 0.00 | **0.94** | .17 | **0.57** | **0.54** | .18 | **0.33** | 0.06 | .08 | 0.04 | .63 |
| dm3 | **0.96** | .19 | **0.52** | **0.67** | .20 | **0.36** | 0.00 | .02 | 0.00 | 0.10 | .09 | 0.05 | .71 |
| pr1 | 0.17 | .20 | 0.08 | 0.00 | .01 | 0.00 | **1.72** | .21 | **0.84** | –0.06 | .17 | –0.03 | .76 |
| pr2 | 0.42 | .25 | 0.22 | **–0.49** | .22 | **–0.25** | **1.62** | .23 | **0.83** | 0.01 | .22 | 0.00 | .74 |
| pr3 | 0.00 | .01 | 0.00 | **0.34** | .18 | **0.18** | **1.47** | .14 | **0.76** | 0.03 | .03 | 0.01 | .73 |
| cm1 | 0.01 | .07 | 0.00 | –0.04 | .07 | –0.03 | 0.01 | .05 | 0.01 | **0.81** | .11 | **0.64** | .39 |
| cm2 | **0.54** | .27 | **0.32** | **–1.00** | .44 | **–0.59** | 0.08 | .21 | 0.05 | **1.56** | .46 | **0.93** | .65 |
| cm3 | 0.01 | .07 | 0.01 | 0.05 | .07 | 0.04 | 0.00 | .05 | 0.00 | **1.17** | .16 | **0.74** | .60 |
| cm4 | 0.00 | .15 | 0.00 | 0.12 | .16 | 0.09 | –0.02 | .09 | –0.01 | **0.93** | .15 | **0.66** | .51 |

*Note*. Statistically significant ($p < .05$) coefficients are in boldface. GK = game knowledge. DM = decision making. PR = pressure. CM = communication.

**Table 4 Study 1, Model 6: Confirmatory Factor Analytic Pattern Coefficients ($\Lambda$), Standard Errors (*SE*), Standardized Pattern Coefficients ($\Lambda^0$), and Percentage of Variance Accounted for ($R^2$)**

| | Factor 1 = GK | | | Factor 2 = DM | | | Factor 3 = PR | | | Factor 4 = CM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\lambda^*_{p1}$ | *SE* | $\lambda^*_{p1}{}^0$ | $\lambda^*_{p2}$ | *SE* | $\lambda^*_{p2}{}^0$ | $\lambda^*_{p3}$ | *SE* | $\lambda^*_{p3}{}^0$ | $\lambda^*_{p4}$ | *SE* | $\lambda^*_{p4}{}^0$ | $R^2$ |
| gk1 | **1.00** | — | **0.47** | — | — | — | — | — | — | **0.61** | 0.15 | **0.33** | .51 |
| gk2 | **2.34** | .42 | **0.84** | — | — | — | — | — | — | — | — | — | .71 |
| gk3 | **1.62** | .25 | **0.74** | — | — | — | — | — | — | — | — | — | .54 |
| dm1 | **0.56** | .17 | **0.25** | **1.00** | — | **0.39** | **0.20** | 0.05 | **0.28** | — | — | — | .54 |
| dm2 | — | — | — | **1.76** | .48 | **0.61** | **0.31** | 0.08 | **0.38** | — | — | — | .65 |
| dm3 | **1.61** | .37 | **0.56** | **1.38** | .35 | **0.41** | — | — | — | — | — | — | .74 |
| pr1 | — | — | — | — | — | — | **1.00** | — | **0.90** | — | — | — | .81 |
| pr2 | — | — | — | — | — | — | **0.68** | .14 | **0.81** | — | — | — | .66 |
| pr3 | — | — | — | **0.63** | .19 | **0.20** | **0.69** | .11 | **0.76** | — | — | — | .71 |
| cm1 | — | — | — | — | — | — | — | — | — | **2.31** | 0.83 | **0.61** | .37 |
| cm2 | **0.71** | .36 | **0.27** | **–1.92** | .91 | **–0.63** | — | — | — | **1.58** | 0.19 | **1.00** | .68 |
| cm3 | — | — | — | — | — | — | — | — | — | **1.35** | 0.21 | **0.77** | .59 |
| cm4 | — | — | — | — | — | — | — | — | — | **0.61** | 0.15 | **0.72** | .51 |

*Note*. Statistically significant ($p < .05$) coefficients are in boldface. GK = game knowledge. DM = decision making. PR = pressure. CM = communication.

***A Priori Measurement Residual Covariances.*** Most (i.e., 4 of 5) of the measurement residual covariances were statistically significant (standardized absolute values ranged from .22 to .26). The covariance between the residual variance of pr3* and the residual variance of cm1*, however, was not statistically significant, $p = .455$, and the standardized value was .09.

***Research Question 3.*** Parameters of interest were the primary (i.e., nonsecondary) pattern coefficients within $\mathbf{\Lambda}$ depicted in Figure 2. Parameter estimates from $\hat{\mathbf{\Lambda}}$ in Model 6 were treated as the population values. This approach was considered a reasonable balance between a priori theory and the more tentative results from the previous research questions.

A modest $N$ (i.e., ~300) provided sufficient power in the vast majority of cases (i.e., 10 of 13 $\lambda$). In three cases (i.e., $\lambda_{\text{dm2*,DM}}$, $\lambda_{\text{dm3*,DM}}$, and $\lambda_{\text{cm2*,CM}}$) a somewhat larger $N$ (i.e., ~450) was needed to provide sufficient power. 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 450 for all parameters of interest is recommended in subsequent studies.

***Potential Sources of Model Misspecification.*** While Model 6 failed to reject the null hypothesis for exact model-data fit, such a failure could be attributed to a lack of power to detect modest deviations from the true model (i.e., a Type II error). Several possible modifications were judged as potentially theoretically defensible and were added to the model depicted in Figure 2 (see * and ** in Table 1). These tentative post hoc modifications were made beginning in Study 2 and a final measurement model was accepted at the conclusion of Study 2.

# Study 2

The fourth research question was investigated in Study 2. New data ($N = 641$) were collected in the US ($n = 404$) and Spain ($n = 237$). Data from Study 1 ($N = 512$) were combined with the new data ($N = 1153$) owing to the importance of group size in multiple-group factorial invariance investigations (Millsap & Yun-Tien, 2004).

## Methods

***Research Question 4.*** The grouping variables were country ($n_{\text{US}} = 729$, $n_{\text{Spain}} = 424$), level of competition ($n_{\text{youth}} = 197$, $n_{\text{high school}} = 460$, $n_{\text{elite}} = 295$), team gender ($n_{\text{male}} = 382$, $n_{\text{female}} = 223$, $n_{\text{mixed}} = 526$), and sport refereed ($n_{\text{basketball}} = 469$, $n_{\text{soccer}} = 270$).[4] In each case, the first group listed within the parentheses of the previous sentence was designated as the reference group. A minimum subgroup sample size of approximately 200 was adopted.

Under the theta parameterization, data were fit to four increasingly restrictive multigroup CFA models for ordinal data for each grouping variable separately (Millsap & Yun-Tien, 2004). Before testing for factorial invariance, the model was imposed separately in each group. Model 1 through Model 4 tested for factorial invariance. Model 1 (baseline) imposed constraints necessary for identification only. Model 2a added the constraint of invariant pattern coefficients (invariant $\mathbf{\Lambda}$) to Model 1. Model 3a added the constraint of invariant thresholds (invariant $\mathbf{\Lambda}$ and $\boldsymbol{\tau}$) to Model 2a. Model 4a added the constraint of an invariant residual covariance matrix (invariant $\mathbf{\Lambda}$, $\boldsymbol{\tau}$, and $\mathbf{\Theta}$) to Model 3a. Nested models were compared as described in Study 1.[5]

If the $\Delta\chi^2_R$ was statistically significant then two additional pieces of information were considered: the guidelines for nested model comparisons, and, the conceptual ramifications of implementing a particular change based on MIs. If a post hoc modification was made the indexing of the Model changed to reflect this (e.g., Model 3a changed to Model 3b if one or more elements within $\tau$ that was constrained to equality by group in Model 3a was freely estimated in at least one group in Model 3b). This approach, while sensitive to capitalization on chance, was viewed as consistent with the preliminary nature of this study.

Following investigation of Research Question 4, the statistical significance of each of the tentatively accepted secondary pattern coefficients and the measurement residual covariances from Study 1 were investigated to determine if there was evidence of replication. This investigation was somewhat confounded with the redundant use of data from Study 1 and all such limits applied. That said, however, over one-half of the data were not from Study 1.

## Results

***Research Question 4.*** Table 5 provides the results of investigations of measurement invariance for each grouping variable not detailed within this section. For textual parsimony only the results by country were detailed below. The model exhibited exact fit both in the US data, $\chi^2_R(30) = 37$, $p = .166$, and in the Spanish data, $\chi^2_R(30) = 23$, $p = .828$.

Model 1 (baseline) exhibited exact fit: $\chi^2_R(69) = 78$, $p = .215$. Model 2a (invariant $\mathbf{\Lambda}$) fit as well as Model 1: $\Delta\chi^2_R(19) = 18$, $p = .521$. Model 3a (invariant $\mathbf{\Lambda}$ and $\boldsymbol{\tau}$) exhibited statistically significant worse fit than Model 2a: $\Delta\chi^2_R(13) = 116$, $p < .001$. A post hoc exploration indicated that five thresholds—gk1$2, gk1$3, cm3$3, gk3$3, and pr3$3—were primarily responsible for the observed non-invariance. More specifically, gk1$2, gk1$3, and cm3$3 were larger in the Spanish data, which indicated that the cumulative frequency of using Categories 1–3 on Item gk1 (and cm3), *understand the basic strategy of the game*, was greater for Spanish referees than for US referees. The opposite interpretation can be made for gk3$3 and pr3$3, as the relevant value for item gk3 and pr3 was greater for US referees. Model 3b adopted both of these post hoc modifications and fit as well as Model 2a: $\Delta\chi^2_R(8) = 7$, $p = .502$.

Model 4a (invariant $\mathbf{\Lambda}$, $\mathbf{\Theta}$, and partially invariant $\boldsymbol{\tau}$) exhibited statistically significant worse fit than Model 3b, $\Delta\chi^2_R(32) = 79$, $p < .001$. A post hoc exploration indicated that three residual variances, dm1*, pr1*, and pr3*, were primarily responsible for the observed non-invariance. More specifically, the residual variance of dm1* was larger in the Spanish data, which indicated that the measurement error associated with responses to dm1, *make critical decisions during competition*, was greater for Spanish referees than for US referees. The opposite interpretation can be made for the residual variance of both pr1* and pr3*, as the measurement error associated with responses to both pr1, *uninfluenced by pressure from players*, and pr3, *uninfluenced by pressure from coaches*, was greater for US referees. Model 4b adopted all three of these post hoc modifications and fit as well as Model 3b: $\Delta\chi^2_R(29) = 38$, $p = .121$. Model 4b imposed invariance for the vast majority (86

## Table 5 Key Results from Study 2, Research Question 4: Measurement Invariance by Level, Team Gender, and Sport Refereed

| | Model–Data Fit | | | | | | Nested Model Comparison | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Group and Model** | **$\chi^2$(*df*)** | **Par** | **RMSEA** | **$CI_{90\%}$** | **CFI** | **TLI** | **Complex** | **$\Delta\chi^2$($\Delta$*df*)** | **$\Delta$RMSEA** | **$\Delta$CFI** | **$\Delta$TLI** |
| Level | | | | | | | | | | | |
| Youth | 38(30) | 87 | .035 | [.000, .067] | .996 | .990 | | | | | |
| High school | 31(30) | 87 | .006 | [.000, .036] | 1.00 | 1.00 | | | | | |
| Elite | 35(30) | 87 | .023 | [.000, .052] | .999 | .996 | | | | | |
| Model 1 | 126(109) | 242 | .021 | [.000, .035] | .998 | .997 | | | | | |
| Model 2a | 182(146)* | 205 | .028 | [.011, .040] | .996 | .994 | Model 1 | 54(37)* | .007 | –.002 | –.003 |
| Model 2b[a] | 148(140) | 211 | .013 | [.000, .030] | .999 | .999 | Model 1 | 27(31) | –.008 | .001 | .002 |
| Model 3a | 213(166)* | 185 | .030 | [.016, .041] | .995 | .993 | Model 2b | 68(26)*** | .017 | –.004 | –.006 |
| Model 3b[b] | 157(157) | 194 | .002 | [.000, .026] | 1.00 | 1.00 | Model 2b | 10(17) | –.011 | .001 | .001 |
| Model 4a | 304(221)*** | 130 | .034 | [.024, .044] | .992 | .991 | Model 3b | 153(64)*** | .032 | –.008 | –.009 |
| Model 4b[c] | 225(212) | 139 | .014 | [.000, .028] | .999 | .999 | Model 3b | 69(55) | .012 | –.001 | –.001 |
| Team Gender | | | | | | | | | | | |
| Male | 35(30) | 87 | .022 | [.000, .046] | .999 | .997 | | | | | |
| Female | 34(30) | 87 | .024 | [.000, .058] | .998 | .995 | | | | | |
| Mixed | 36(30) | 87 | .020 | [.000, .040] | .999 | .997 | | | | | |
| Model 1 | 136(109) | 242 | .026 | [.008, .039] | .998 | .995 | | | | | |
| Model 2a | 173(146) | 205 | .022 | [.000, .034] | .998 | .996 | Model 1 | 48(37) | –.004 | .000 | .001 |
| Model 3a | 228(172)** | 179 | .029 | [.018, .038] | .995 | .994 | Model 2a | 60(26)*** | .007 | –.003 | –.002 |
| Model 3b[d] | 185(167) | 184 | .017 | [.000, .030] | .998 | .998 | Model 2a | 11(21) | –.005 | .000 | .002 |
| Model 4a | 308(230)*** | 121 | .030 | [.020, .038] | .993 | .993 | Model 3b | 129(63)*** | .013 | –.005 | –.005 |
| Model 4b[e] | 254(226) | 125 | .018 | [.000, .029] | .998 | .998 | Model 3b | 72(59) | .001 | .000 | .000 |

*(continued)*

Table 5 *(continued)*

| Group and Model | Model–Data Fit |  |  |  |  |  | Nested Model Comparison |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  | $\chi^2(df)$ | Par | RMSEA | $CI_{90\%}$ | CFI | TLI | Complex | $\Delta\chi^2(\Delta df)$ | ΔRMSEA | ΔCFI | ΔTLI |
| Sport |  |  |  |  |  |  |  |  |  |  |  |
| Basketball | 30(30) | 87 | .000 | [.000, .035] | 1.00 | 1.00 |  |  |  |  |  |
| Soccer | 23(30) | 87 | .000 | [.000, .030] | 1.00 | 1.00 |  |  |  |  |  |
| Model 1 | 74(69) | 165 | .013 | [.000, .034] | .999 | .999 |  |  |  |  |  |
| Model 2a | 149(88)*** | 146 | .043 | [.031, .055] | .991 | .985 | Model 1 | 62(19)*** | .030 | –.008 | –.014 |
| Model 2b[f] | 92(80) | 154 | .020 | [.000, .037] | .998 | .997 | Model 1 | 17(11) | .007 | –.001 | –.002 |
| Model 3a | 224(94)*** | 140 | .061 | [.051, .072] | .982 | .970 | Model 2b | 129(14)*** | .041 | –.016 | –.027 |
| Model 3b[g] | 97(87) | 147 | .018 | [.000, .035] | .999 | .997 | Model 2b | 5(7) | –.002 | .001 | .000 |
| Model 4a | 196(119)*** | 115 | .042 | [.031, .052] | .989 | .986 | Model 3b | 107(32)*** | .024 | –.010 | –.011 |
| Model 4b[h] | 130(112) | 122 | .021 | [.000, .035] | .998 | .997 | Model 3b | 34(25) | .003 | –.001 | .000 |

*Note.* Model 1 = Baseline (i.e., identification constraints only). Model 2a = Invariant pattern coefficients (**Λ**). Model 3a = Invariant **Λ** and thresholds (**τ**). Model 4 = Invariant **Λ**, **τ**, and unique factor (co)variances (**Θ**).

[a] Non-invariant **Λ**: gk3*,GK (smaller for elite); dm1*,GK (largest for youth; smallest for elite); dm1*,PR (larger for elite); dm3*,PR, cm4*,CM (larger for high school).

[b] Non-invariant **τ**: dm2$3, dm3$3, pr3$3 (smallest for youth, largest for elite); gk1$2 (smaller for youth); gk3$2 (larger for high school); gk3$3 (larger for elite).

[c] Non-invariant **Θ**: gk2* (larger for high school); cm2* (larger for youth); cm3*, pr1* with cm3* (smaller for elite); cm2* with cm4* (larger for elite); pr3*, pr2* with dm3*, pr3* with dm2* (smaller for youth).

[d] Non-invariant **τ**: gk1$3 (smaller for mixed); ps3$3, gk3$3 (larger for mixed); cc4$2, dm1$3 (larger for female).

[e] Non-invariant **Θ**: pr3* (smaller for male); gk1* (larger for female); pr1*, dm1* (smaller for mixed).

[f] Non-invariant **Λ**: dm1*,GK, gk2*,GK, dm1*,PR, cm1*,CM, cm3*,CM, cm4*,CM (smaller for basketball); pr3*,PR (smaller for soccer).

[g] Non-invariant **τ**: gk1$3, gk2$1 gk3$1, cm3$1, cm4$3 (larger for soccer); gk3$3, pr3$3 (larger for basketball).

[h] Non-invariant **Θ**: cm2*, cm4*, dm1* with cm4*, pr2 with gk2 (larger for soccer); gk2*, gk3*, pr2* (larger for basketball).

*$p < .05$. **$p < .01$. ***$p < .001$.

of 94) of measurement parameters (i.e., 23 of 23 $\lambda$, 34 of 39 $\tau$, and 29 of 32 $\theta$) by country and exhibited exact fit, $\chi^2_R(125) = 137$, $p = .213$.

Results by level of competition, team gender, and sport refereed are only briefly summarized due to space limitations. The accepted model with regard to level of competition, Model 4b, imposed invariance for the majority (75 of 94) of measurement parameters (i.e., 18 of 23 $\lambda$, 33 of 39 $\tau$, and 24 of 32 $\theta$) and exhibited exact fit, $\chi^2_R(212) = 225$, $p = .254$. The accepted model with regard to team gender, Model 4b, imposed invariance for the vast majority (85 of 94) of measurement parameters (i.e., 23 of 23 $\lambda$, 34 of 39 $\tau$, and 28 of 32 $\theta$) and exhibited exact fit, $\chi^2_R(226) = 254$, $p = .119$. The accepted model with regard to sport refereed, Model 4b, imposed invariance for the majority (73 of 94) of measurement parameters (i.e., 16 of 23 $\lambda$, 32 of 39 $\tau$, and 25 of 32 $\theta$) and exhibited exact fit, $\chi^2_R(226) = 254$, $p = .119$. A reasonable answer to Research Question 4 was that there was strong evidence for partial factorial invariance by country, level of competition, team gender, and sport refereed. Degree of non-invariance appeared to be greatest for level of competition and sport refereed (elaborated upon in the Discussion).

***Accepted Measurement Model.*** Information from Study 2, regarding the secondary pattern coefficients and measurement residual covariances in particular, informed the accepted measurement model. Most (i.e., 9 of 10) of the secondary pattern coefficients generally were statistically significant in Model 4b across the four multigroup analyses. Most (i.e., 16 of 19) of the measurement residual covariances specified generally were statistically significant in Model 4b across the four multigroup analyses. Figure 3 depicts the accepted measurement model.

# Study 3

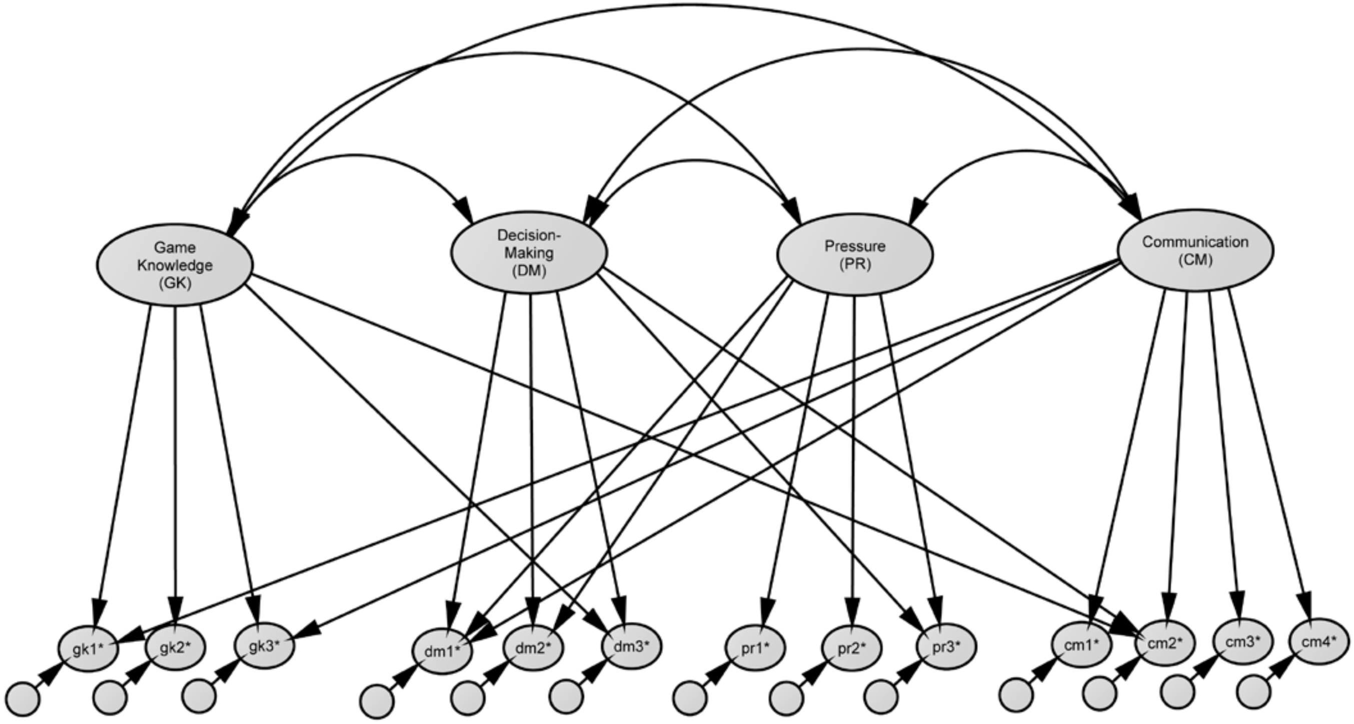
The fifth research question was investigated in Study 3. New data ($N = 456$) were collected in the United States ($n = 249$) and Spain ($n = 207$). In a subset of the Spanish data that were collected in-person ($n = 159$), responses to the SSCQ were collected.[6]

## Methods

***Research Question 5.*** The correlation matrix between the dimensions of referee self-efficacy and the eight potential sources of efficacy information (i.e., years of referee experience, highest level refereed, and the six SSCQ dimensions) was estimated in Model 1. Correlations between the predictors that were not statistically significant were not specified in Model 2 to reduce the number of parameters estimated. In Model 2, the eight potential sources of efficacy information were specified as predictors of each dimension of referee self-efficacy within a multivariate regression with latent outcomes. Magnitude of the variance jointly accounted for in each dimension of referee self-efficacy was assessed using Cohen's (1988) effect size heuristics.

## Results

***Research Question 5.*** Table 6 displays the correlation matrix from Model 1. Each potential source of efficacy information, except for social support, had a statistically



Measurment residual covariances
gk1* with pr3*, cm2*, cm3*
gk2* with cm2,* cm4*, dm3*, pr2*
gk3* with pr3*

Measurment residual covariances
dm1* with cm4*, and pr1*
dm2* with cm4* and pr3*
dm3* with pr2*

Measurment residual covariances
pr1* with cm3*
pr3* with cm1*

Measurment residual covariances
cm1* with cm3*

**Figure 3** — Accepted measurement theory for the REFS. Observed variables were omitted to reduce clutter.

## Table 6 Descriptive Statistics from Model 1 in Study 3: Correlations, Means, and Standard Deviations

| | Dimensions of Referee Self-Efficacy | | | | Proposed Sources of Referee Self-Efficacy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GK | DM | PR | CM | YE | HL | SS | PMP | EC | SF | PA | VE |
| Game knowledge (GK) | 1.00 | | | | | | | | | | | |
| Decision making (DM) | .80*** | 1.00 | | | | | | | | | | |
| Pressure (PR) | .63*** | .77*** | 1.00 | | | | | | | | | |
| Communication (CM) | .66*** | .85*** | .67*** | 1.00 | | | | | | | | |
| Years experience (YE) | .33*** | .42*** | .33*** | .27*** | 1.00 | | | | | | | |
| Highest level (HL) | .37*** | .40*** | .27*** | .22*** | .51*** | 1.00 | | | | | | |
| Social support (SS) | .16 | .14 | .14 | .15 | –.07 | –.13 | 1.00 | | | | | |
| Physical/mental preparation (PMP) | .48*** | .35*** | .40*** | .21** | .14 | .16* | .46*** | 1.00 | | | | |
| Environmental comfort (EC) | .13 | .29*** | .25** | .23** | .03 | –.04 | .48*** | .32*** | 1.00 | | | |
| Situational favorableness (SF) | .15 | .28*** | .21** | .23** | .17* | .08 | .61*** | .53*** | .65*** | 1.00 | | |
| Past accomplishments (PA) | .20* | .23** | .21** | .10 | .02 | –.03 | .48*** | .45*** | .45*** | .49*** | 1.00 | |
| Vicarious experience (VE) | .24** | .24** | .21** | .09 | .15 | .08 | .45*** | .49*** | .39*** | .63*** | .28*** | 1.00 |
| *M* | 0.00 | 0.00 | 0.00 | 0.00 | 11.29 | 2.08 | 4.87 | 5.34 | 4.66 | 4.59 | 5.22 | 4.41 |
| *SD* | 0.52 | 1.62 | 1.29 | 0.82 | 10.14 | 0.80 | 1.15 | 0.90 | 1.32 | 1.02 | 1.13 | 1.42 |

*$p < .05$. **$p < .01$. ***$p < .001$.

significant positive correlation with at least one dimension of referee self-efficacy. Absolute values of the correlations between the predictors ranged from .02 to .65.

Model 2 exhibited exact fit, $\chi^2_R(117) = 133$, $p = .144$. Table 7 displays the key results from Model 2. Potential sources of referee self-efficacy information combined to account for a moderate or large amount of variance in each dimension of referee self-efficacy $R^2$ ranged from $R^2_{CM} = 16.1\%$ to $R^2_{GK} = 44.7\%$). Years of referee experience, highest level refereed, physical/mental preparation, and environmental comfort each exerted at least one statistically significant positive direct effect. Years of referee experience and physical/mental preparation both exerted a statistically significant direct effect on each dimension of referee self-efficacy. Highest level refereed exerted a statistically significant direct effect on both game knowledge efficacy and decision-making efficacy. Environmental comfort exerted a statistically significant direct effect on both pressure efficacy and communication efficacy.

# Discussion

The purpose of this multistudy report was to develop, and then to provide initial validity evidence for measures derived from, the REFS. Development of the REFS was congruent with self-efficacy theory and research in sport, was guided by content and measurement experts, and yielded a practical instrument to collect data from an important population in sport that has largely been ignored. Preliminary evidence for the internal and external validity of measures derived from the REFS was provided. There are limits, however, to the evidence provided.

In Study 1, validity evidence was provided for the more restrictive CFA measurement model depicted in Figure 2 versus the more flexible ESEM measurement model depicted in Figure 1. Further, in both Study 1 and Study 2, validity evidence was provided for at least close fit of the measurement model. Thus, it appears that the measurement model for the REFS may be fairly well understood and that a CFA approach may be preferred over an ESEM approach in future research with the REFS. When sufficient a priori measurement theory exists, CFA generally is preferred over ESEM because ". . . it is better to tell a statistical program what the true theoretical model is and then to receive confirmatory feedback from the program than it is to tacitly ask an inanimate statistical program to co-develop a theoretical model" (Myers, Chase, et al., 2011, p. 799).[7] The accepted measurement model depicted in Figure 3 appears to be adequately powered with a moderate sample size (e.g., a minimum of $N$ ~300 with a desired $N \geq 450$), which should make the REFS feasible to implement in many study designs in the future.

An assumption in the development of the REFS was that the intended population for the REFS spans several potentially important grouping variables. In Study 2, evidence was provided for partial factorial invariance by country, level of competition refereed, team gender, and sport refereed. Empirical implications of partial factorial invariance versus full factorial invariance, particularly for ordinal data, are largely unclear (Millsap & Kwok, 2004). The degree of non-invariance observed in this study (i.e., at least 78% of measurement parameters were specified as invariant by each grouping variable) was assumed to exert marginal practical impact on the ability to compare measures across the identified grouping variables. The extent to which this assumption is reasonable was a potential limitation of the current study. At this point, however, a reasonable conclusion is that measures derived from the REFS appear to

**Table 7 Key Results from Model 2 in Study 3, Research Question 5: Potential Sources of Referee Self-Efficacy Information**

| | Game Knowledge | | | Decision Making | | | Pressure | | | Communication | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | $\gamma$ | *SE* | *p* | $\gamma$ | *SE* | *p* | $\gamma$ | *SE* | *p* | $\gamma$ | *SE* | *p* |
| Years of referee experience | **0.14** | .06 | .029 | **0.51** | .18 | .005 | **0.41** | .13 | .001 | **0.15** | .07 | .028 |
| Highest level | **0.17** | .07 | .014 | **0.53** | .21 | .012 | 0.18 | .14 | .206 | 0.14 | .09 | .099 |
| Social support | 0.00 | .05 | .908 | –0.19 | .18 | .284 | –0.09 | .12 | .437 | 0.02 | .08 | .823 |
| Physical/mental preparation | **0.32** | .09 | .000 | **0.57** | .24 | .016 | **0.63** | .15 | .000 | **0.17** | .09 | .047 |
| Environmental comfort | 0.05 | .05 | .343 | 0.33 | .19 | .081 | **0.27** | .11 | .018 | **0.12** | .06 | .049 |
| Situational favorableness | –0.19 | .13 | .131 | –0.23 | .31 | .463 | –0.34 | .23 | .137 | 0.04 | .13 | .742 |
| Past accomplishments | 0.02 | .04 | .631 | 0.10 | .18 | .562 | 0.04 | .10 | .703 | –0.06 | .07 | .391 |
| Vicarious experience | 0.05 | .05 | .295 | 0.13 | .13 | .334 | 0.07 | .09 | .459 | –0.06 | .07 | .418 |
| Percentage of variance accounted for | 44.7% | | | 40.8% | | | 33.4% | | | 16.1% | | |

*Note*. Statistically significant ($p < .05$) coefficients are in boldface.

be relatively comparable across the grouping variables studied. Future research that explores why the identified measurement parameters may be non-invariant would be useful—particularly if each relevant item can be revised in a relatively minor way to encourage invariance across the various subgroups within the intended population for the REFS. Another option would be to develop separate REFS instruments for various subgroups, particularly by level of competition refereed and/or sport refereed, as has been done to some degree in related research (e.g., Myers et al., 2010). If separate REFS instruments for various subgroups are put forth (perhaps using some REFS items as common items across forms) such an effort may be well served by forming subgroup-specific focus groups. That said, and as can be viewed in Table 5, it appears that the REFS exhibits at least close model–data fit within each of the studied subgroups individually. For example, the first two rows under "Sport" in Table 5 suggest that the "Basketball" and "Soccer" subgroups each passed the exact fit test when fit to the data separately. Thus, the primary model–data fit challenge appears to be that the magnitude of some measurement parameters may not be homogenous across some subgroups, and thus, the comparability of the measures across these subgroups may be compromised if such comparisons are made.

In terms of sources of referee self-efficacy beliefs, it is not surprising that years of experience and physical/mental preparation were predictive of all four factors of referee self-efficacy. These sources are based on one's mastery experiences and sense of readiness, which are considered within self-efficacy theory as the most dependable for forming efficacy judgments. The highest level that one has officiated also is based on mastery because one needs to have experience to move up in officiating levels. But this source was a significant predictor only for decision-making and game-knowledge efficacy. Making critical decisions and having an understanding of the strategy of the game may become more essential skills as one gets to higher levels of officiating and, as referees have more experience at higher levels, they may become more confident in the decision-making and game-knowledge aspects of their performance.

Environmental comfort (e.g., with the venue) was a significant predictor of efficacy beliefs to be unaffected by pressure from spectators, coaches, and players as well as communication efficacy. These may be venues where there is more crowd control, where spectators are further away from the referees (therefore, they can communicate more effectively with other referees, coaches, and players), and in locations where they know the crowds behave respectfully. If crowds are behaving respectfully, referees can have more confidence to approach coaches and players to communicate with them. Referees may also feel like they can focus more on their job and less on various external pressures when they perform in a comfortable venue.

Four additional sources of information were not significant predictors of any of the dimensions of referee self-efficacy: past performance accomplishments, perceived support, vicarious experience, and situational favorableness. Unlike coaching efficacy (Feltz et al., 1999), perceived social support was not an important source of referee self-efficacy information. It may be that the controversial nature of officiating makes social support from others to be a relatively unreliable source of efficacy for referees. Likewise, referees may spend little time vicariously experiencing other officials' successful performance, especially in the face of their own mastery experience. Lastly, situational favorableness, which includes favorable weather, co-official, and assigned game/match, is out of the control of officials. Vealey et al. (1998) considered this source as an uncontrollable source of efficacy information, and thus, it is not surprising that referees would not consider this to be a dependable source of efficacy information. It was surprising, however, that the past performance accomplishments subscale of the SSCQ did not predict any dimension of referee self-efficacy. We speculate that this subscale of the SSCQ may need to be rethought, at least with regard to referee self-efficacy, and this speculation is consistent with the statistically nonsignificant bivariate correlations between this subscale and other indicators of mastery experience (i.e., years of officiating experience and highest level refereed—see Table 6).

Primary limits for evidence provided in this multistudy report included weaknesses of a model-generating approach, a small sample of female referees, omission of possible outcomes of referee self-efficacy, and that the current conceptualization of referee efficacy is but one of many possibilities. A model-generating approach was employed in several instances in this investigation (e.g., post hoc modifications). While model generation is appropriate when there is insufficient previous research to support a strictly disconfirmatory approach, such an approach is susceptible to capitalizing on chance and subsequent results should be viewed with caution before replication (Jöreskog, 1993). Another limitation of the current investigation was the small sample of females ($n$ = 115). The underrepresentation of females in the current study, although unfortunate, was consistent with established trends in officiating (Casey, 1992; Muster, 2001). Future research with a larger sample of female referees is warranted, in part, to determine to what degree that measures derived from the REFS are comparable across referee gender. Another limitation was that potential outcomes of referee self-efficacy were not a focus of the current investigation. Future research that investigates the ability of measures derived from the REFS to predict potential outcomes of referee self-efficacy, including those proposed by Guillén and Feltz (2011)—referee/athlete/coach behavior, referee satisfaction, referee performance, referee stress, and athlete rule violations—could make an important contribution to the literature, particularly if done simultaneously with proposed sources of referee efficacy (e.g., exploring dimensions of referee efficacy as mediators). Finally, even though evidence was provided for the conceptualization of referee self-efficacy implemented in the REFS, the conceptualization put forward should be viewed as only one of many possibilities. Competing conceptualizations of referee self-efficacy that may include additional dimensions (e.g., physical fitness efficacy) and/or explore the utility of subgroup specific REFS instruments should be encouraged and tested.

There is much work yet to do to investigate the full utility of the referee self-efficacy construct. Given the framework provided by Guillén and Feltz (2011) and the results of the current study, however, a population-specific conceptual and measurement model for referee self-efficacy now has been provided and may serve as a guide to subsequent advances in the relevant research. A fruitful area of future research may be to investigate theory-based alternate forms of each relationship (e.g., nonlinear and/or interactions by sport refereed) proposed in the preliminary conceptual model put forth by Guillén and Feltz.

## Notes

1. Guillén and Feltz (2011) succinctly referred to referee self-efficacy as *refficacy*. In an effort to keep the relevant task (i.e., refereeing) and the relevant construct (i.e., self-efficacy) separate, however, we do not adopt the “refficacy” expression in this manuscript.

2. The aim of the 2-hr qualitative focus group in Guillén and Feltz (2011) was "not to conduct a research study but rather to seek input in developing our model" (p. 2). Two members of the expert group in the current study also were involved with the focus group in Guillén and Feltz.

3. It should also be noted that a Type II error may occur with greater frequency than is generally judged to be tolerable when both models are misspecified and sample size is small (Yuan & Bentler, 2004). As can be viewed in Table 2, the only case where a Type II error was possible with regard to Research Questions 1 and 2 was when Model 6 was compared with Model 4.

4. Data from Division I university and/or semiprofessional referees ($n$ = 255), and professional and/or international referees ($n$ = 40) formed the elite group ($n$ = 295). Remaining data from sport refereed ($n$ = 414) was distributed across several sports in relatively small amounts.

5. When the more complex model is correctly specified but the simpler model is misspecified, a Type II error may occur with greater frequency than is generally judged to be tolerable, particularly when the sample size is small (Yuan & Bentler, 2004).

6. A limitation in the design of Study 3 (e.g., a small subsample of referees completed the SSCQ) did not allow for a rigorous investigation of the assumption of invariance of the direct effects of potential sources of referee self-efficacy on dimensions of referee self-efficacy by possible moderating variables (e.g., sport refereed). Future research in this area may be warranted.

7. In some cases, particularly when little is known a priori, even the ESEM framework may be viewed as too restrictive and automated causal discovery techniques may be viewed by some scholars as more appropriate (e.g., Landsheer, 2010; Rozeboom, 2008).

## References

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

Asparouhov, T., & Muthén, B.O. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397–438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B.O. (2010). *Simple second order chi-square correction*. Retrieved from Mplus website: http://www.statmodel.com/download/WLSMV_new_chi21.pdf

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150. doi:10.1207/S15327906MBR3601_05

Casey, A. (1992). Title IX and women officials—How have they been affected? *Journal of Physical Education, Recreation & Dance, 63,* 45–47.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Feltz, D.L. (1982). A path analysis of the causal elements in Bandura's theory of self-efficacy and an anxiety-based model of avoidance behavior. *Journal of Personality and Social Psychology, 42,* 764–781. doi:10.1037/0022-3514.42.4.764

Feltz, D.L., & Chase, M.A. (1998). The measurement of self-efficacy and confidence in sport. In J.L. Duda (Ed.), *Advancements in sport and exercise psychology measurement* (pp. 65–80). Morgantown, WV: Fitness Information Technology.

Feltz, D.L., Chase, M.A., Moritz, S.E., & Sullivan, P.J. (1999). A conceptual model of coaching efficacy: Preliminary investigation and instrument development. *Journal of Educational Psychology, 91,* 765–776. doi:10.1037/0022-0663.91.4.765

Feltz, D.L., & Lirgg, C.D. (1998). Perceived team and player efficacy in hockey. *The Journal of Applied Psychology, 83,* 557–564. PubMed doi:10.1037/0021-9010.83.4.557

Feltz, D.L., Short, S.E., & Sullivan, P.J. (2008). *Self-efficacy in sport*. Champaign, IL: Human Kinetics.

Finney, S.J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In R.C. Serlin (Series Ed.) & G.R. Hancock & R.O. Mueller (Vol. Eds.), *Structural equation modeling: A second course* (pp. 269–313). Greenwich, CT: Information Age.

Guillén, F., & Feltz, D.L. (2011). A conceptual model of referee efficacy. *Frontiers in Psychology, 2,* 25. PubMed doi:10.3389/fpsyg.2011.00025

Hancock, G.R., & Mueller, R.O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S.H.C. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Past and present. A festschrift in honor of Karl G. Jöreskog* (pp. 195–261). Chicago: Scientific Software International, Inc.

Hayashi, K., Bentler, P.M., & Yuan, K. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling, 14,* 505–526. doi:10.1080/10705510701301891

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi:10.1080/10705519909540118

Jackson, B., Beauchamp, M.R., & Knapp, P. (2007). Relational efficacy beliefs in athlete dyads: An investigation using actor-partner interdependence models. *Journal of Sport & Exercise Psychology, 29,* 170–189 PubMed. PubMed

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

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 467–482. doi:10.1207/s15327906mbr2301_4

Landsheer, J.A. (2010). The specification of causal models with Tetrad IV: a review. *Structural Equation Modeling, 17*(4), 703–711. doi:10.1080/10705511.2010.510074

Lent, R.W., & Lopez, F.G. (2002). Cognitive ties that bind: A tripartite view of efficacy beliefs in growth-promoting relationships. *Journal of Social and Clinical Psychology, 21,* 256–286. doi:10.1521/jscp.21.3.256.22535

MacCallum, R.C., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504. PubMed doi:10.1037/0033-2909.111.3.490

Marsh, H.W., Hau, K-T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220. doi:10.1207/s15327906mbr3302_1

Marsh, H.W., Hau, K.T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. doi:10.1207/s15328007sem1103_2

Marsh, H.W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A.J.S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22,* 471–491. PubMed doi:10.1037/a0019227

Millsap, R.E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Millsap, R.E., & Kwok, O-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93–115. PubMed doi:10.1037/1082-989X.9.1.93

Millsap, R.E., & Yun-Tien, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515. doi:10.1207/S15327906MBR3903_4

Moritz, S.E., Feltz, D.L., Fahrbach, K.R., & Mack, D.E. (2000). The relation of self-efficacy measures to sport performance: a meta-analytic review. *Research Quarterly for Exercise and Sport, 71,* 280–294 PubMed. PubMed

Muster, B. (2001). Minority hiring practices in professional sports. *The Sports Journal, 4*. Retrieved October 5, 2011 from http://www.thesportjournal.org/tags/volume-4-number-4.

Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indictors. *Psychometrika, 49,* 115–132. doi:10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L.K., & Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620. doi:10.1207/S15328007SEM0904_8

Myers, N.D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. *Research Quarterly for Exercise and Sport, 82,* 412–423. PubMed

Myers, N.D., Chase, M.A., Beauchamp, M.R., & Jackson, B. (2010). The Coaching Competency Scale II – High School Teams. *Educational and Psychological Measurement, 70,* 477–494. doi:10.1177/0013164409344520

Myers, N.D., Chase, M.A., Pierce, S.W., & Martin, E. (2011). Coaching efficacy and exploratory structural equation modeling: A substantive-methodological synergy. *Journal of Sport & Exercise Psychology, 33,* 779–806. PubMed

Myers, N.D., Feltz, D.L., Chase, M.A., Reckase, M.D., & Hancock, G.R. (2008). The Coaching Efficacy Scale II - High School Teams. *Educational and Psychological Measurement, 68,* 1059–1076. doi:10.1177/0013164408318773

Myers, N.D., Feltz, D.L., & Wolfe, E.W. (2008). A confirmatory study of rating scale category effectiveness for the coaching efficacy scale. *Research Quarterly for Exercise and Sport, 79,* 300–311. PubMed doi:10.5641/193250308X13086832905752

Myers, N.D., Wolfe, E.W., & Feltz, D.L. (2005). An evaluation of the psychometric properties of the coaching efficacy scale for American coaches. *Measurement in Physical Education and Exercise Science, 9,* 135–160. doi:10.1207/s15327841mpee0903_1

Myers, N.D., Wolfe, E.W., Feltz, D.L., & Penfield, R.D. (2006). Identifying differential item functioning of rating scale items with the Rasch model: An introduction and an application. *Measurement in Physical Education and Exercise Science, 10,* 215–240. doi:10.1207/s15327841mpee1004_1

Rozeboom, Wm.W. (2008). The problematic importance of hypotheses. *Journal of Clinical Psychology (Savannah, Ga.), 64*(9), 1109–1127.

Saris, W.E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16,* 561–582. doi:10.1080/10705510903203433

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. PubMed doi:10.1037/1082-989X.7.2.147

Short, S.E., Sullivan, P.S., & Feltz, D.L. (2005). Development and preliminary validation of the collective efficacy questionnaire for sports. *Measurement in Physical Education and Exercise Science, 9,* 181–202. doi:10.1207/s15327841mpee0903_3

Spink, K.S. (1990a). Group cohesion and collective efficacy of volleyball teams. *Journal of Sport & Exercise Psychology, 12,* 301–311.

Spink, K.S. (1990b). Collective efficacy in the sport setting. *International Journal of Sport Psychology, 21,* 380–395.

Tuero, C., Tabernero, B., Marquez, S., & Guillen, F. (2002). Análisis de los factores que influyen en la práctica del arbitraje [Analysis of the factors affecting the practice of refereeing]. *SCAPE, 1*(1), 7–16.

Vealey, R.S., Hayashi, S.W., Garner-Holman, G., & Giacobbi, P. (1998). Sources of Sport confidence: conceptualization and instrument development. *Journal of Sport & Exercise Psychology, 20,* 54–80.

Yuan, K.H., & Bentler, P.M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64,* 737–757. doi:10.1177/0013164404264853

*Manuscript submitted: December 16, 2011*
*Revision accepted: August 16, 2012*