Identifying Differential Item Functioning of Rating Scale Items With the Rasch Model: An Introduction and an Application

Article *in* Measurement in Physical Education and Exercise Science · December 2006

DOI: 10.1207/s15327841mpee1004_1

CITATIONS
38

READS
723

4 authors, including:

Nicholas D. Myers
Michigan State University
119 PUBLICATIONS **2,347** CITATIONS
SEE PROFILE

Edward W Wolfe
Pearson
105 PUBLICATIONS **3,061** CITATIONS
SEE PROFILE

Deborah Feltz
Michigan State University
228 PUBLICATIONS **9,235** CITATIONS
SEE PROFILE

MEASUREMENT IN PHYSICAL EDUCATION AND EXERCISE SCIENCE, *10*(4), 215–240

Identifying Differential Item Functioning of Rating Scale Items With the Rasch Model: An Introduction and an Application

Nicholas D. Myers
Department of Educational and Psychological Studies
University of Miami

Edward W. Wolfe
Department of Educational Leadership and Policy Studies
Virginia Polytechnic Institute and State University

Deborah L. Feltz
Department of Kinesiology
Michigan State University

Randall D. Penfield
Department of Educational and Psychological Studies
University of Miami

This study (a) provided a conceptual introduction to differential item functioning (DIF), (b) introduced the multifaceted Rasch rating scale model (MRSM) and an associated statistical procedure for identifying DIF in rating scale items, and (c) applied this procedure to previously collected data from American coaches who responded to the coaching efficacy scale (CES; Feltz, Chase, Moritz, & Sullivan, 1999). In this study, an item displayed DIF if coaches from different groups were more or less likely to endorse that item once coaches were matched on the efficacy of interest, where Motivation, Game Strategy, Technique, and Character Building efficacies defined coaching efficacy. Coach gender and level coached were selected

Correspondence should be addressed to Nicholas D. Myers, School of Education, University of Miami, 311E Merrick Building, P.O. Box 248065, Coral Gables, FL 33124–2040. E-mail: nmyers@miami.edu

as the grouping variables. None of the Technique and Character Building items exhibited DIF based on coach gender or level coached. One of the Motivation items and one of the Game Strategy items exhibited DIF based on coach gender. Two of the Motivation items exhibited DIF based on level coached.

Key words: coaching efficacy, differential item functioning, Rasch model, item response theory

Messick (1989) described the generalizability of measures as being an essential aspect of measurement validity. Specifically, he suggested that, for measures to be interpretable across a variety of contexts, instruments, or population subgroups, the measures must carry the same meaning across these aspects of the measurement situation. If measures mean one thing for one group but mean something different for another group, it is not possible to make statements concerning relative levels of performance between the two groups. It has become common practice in many fields to evaluate the degree to which the meaningfulness of measures generalizes across subgroups within a population. Studies that focus on this aspect of validity at the item-level within the instrument are referred to as studies of differential item functioning, that is, DIF (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

"An item displays DIF if the examinees from different groups have different probabilities or likelihoods of success on the item after conditioning or matching on the ability the item is intended to measure" (Clauser & Mazor, 1998, p. 268). From here, we backtrack to introduce DIF analysis for rating scale items (i.e., polytomous), identified with a specific type of statistical procedure associated with the Rasch model (a latent variable measurement model), before applying what we will have introduced. A working understanding of item response theory (IRT), specifically the Rasch model, is assumed.

Neither of the introductions that follow should be construed as original thinking by the authors. Rather, the introductions are based on the authors' understanding of earlier work focused on extending DIF procedures for dichotomous data to polytomous data (Muraki, 1993; Rogers & Swaminathan, 1993; Welch & Hoover, 1993) and subsequent syntheses of this work in the education literature (Clauser & Mazor, 1998; Penfield & Lam, 2000; Potenza & Dorans, 1995). The rationale for providing these introductions is to put an aspect of validity evidence, DIF, and a common statistical procedure for assessing DIF, into familiar and accessible contexts for the *Measurement in Physical Education and Exercise Science* readership. Identifying DIF within these contexts could have important applications in physical education and exercise science because constructs are routinely measured with rating scale items across substantively interesting groups of subjects, and the mod-

els used to create measures from the participants' responses to these items frequently assume that these groupings exert no impact on the interactions between the persons and the items. Therefore, the purposes of this study were to (a) provide a conceptual introduction to DIF, (b) introduce the multifaceted Rasch rating scale model (MRSM) and an associated statistical procedure for identifying DIF in rating scale items, and (c) apply this procedure to previously collected data from American coaches who responded to the Coaching Efficacy Scale (CES; Feltz, Chase, Moritz, & Sullivan, 1999).

Prior to introducing the DIF procedures, we describe the CES because it will serve as the basis for our examples. Coaching efficacy is the extent to which a coach believes that he or she has the capacity to affect the learning and performance of his or her athletes (Feltz et al., 1999). The CES is the only published instrument purported to measure coaching efficacy. Important initial steps in the validation process of the CES already reported include the explication of a conceptual framework for scores derived from the instrument (Myers, Wolfe, & Feltz, 2005), explaining how the construct is operationalized within the instrument (Myers, Wolfe, et al., 2005), revealing the instrument's development process (Feltz et al., 1999), generating an internal model of the construct (Feltz et al., 1999; Lee, Malete, & Feltz, 2002; Myers, Wolfe, et al., 2005), and predicting how measures of the construct relate to other variables—an external model (Feltz et al., 1999; Malete & Feltz, 2000; Myers, Vargas-Tonsing, & Feltz, 2005; Sullivan & Kent, 2003; Vargas-Tonsing, Warners, & Feltz, 2003). Of particular importance in this study was evidence for the internal model (i.e., components of coaching efficacy posited to exert influence on responses to the CES items).

The internal model for the CES is illustrated in Figure 1 (see the Appendix for item content). This model posits that four specific efficacies (i.e., Motivation [ME], Game Strategy [GS], Technique [TE], and Character Building [CB]) are related to one another and define coaching efficacy. ME is specified to influence responses to seven items and is defined as the confidence coaches have in their abilities to affect the psychological mood and psychological skills of athletes. GS is specified to influence responses to seven items and is defined as the confidence coaches have in their abilities to lead during competition. TE is specified to influence responses to six items and is defined as the belief coaches have in their instructional and diagnostic skills. CB is specified to influence responses to four items and is defined as the confidence coaches have in their abilities to influence the personal development and positive attitude toward sport in athletes. In the internal model, subgroupings of coaches (e.g., coach's gender or level coached) are not specified to influence responses to the CES items. It is not incongruent with Horn's (2002) model of coaching effectiveness, however, that coaches' personal characteristics may influence their beliefs (see Figure 2) and therefore responses to items measuring those beliefs (e.g., ME items) even after controlling for the relevant ability level (e.g., ME).

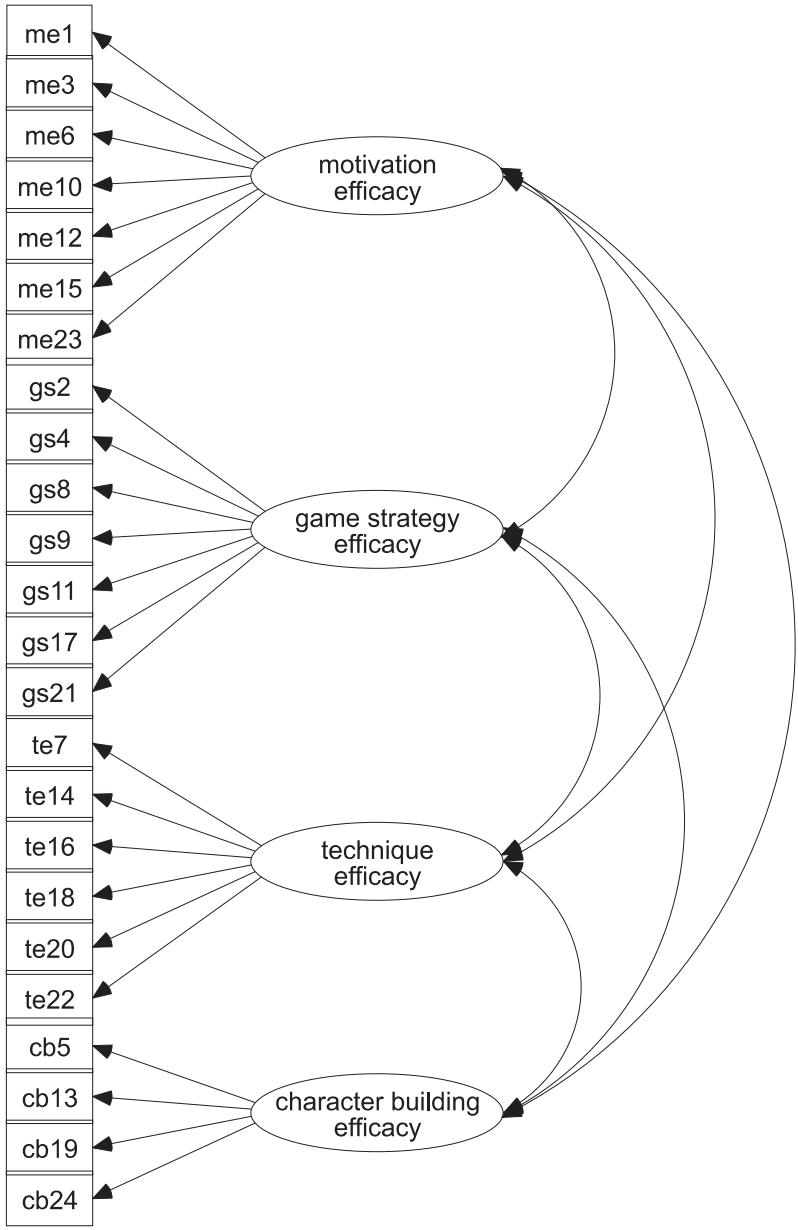

FIGURE 1 Multidimensional model of the Coaching Efficacy Scale.

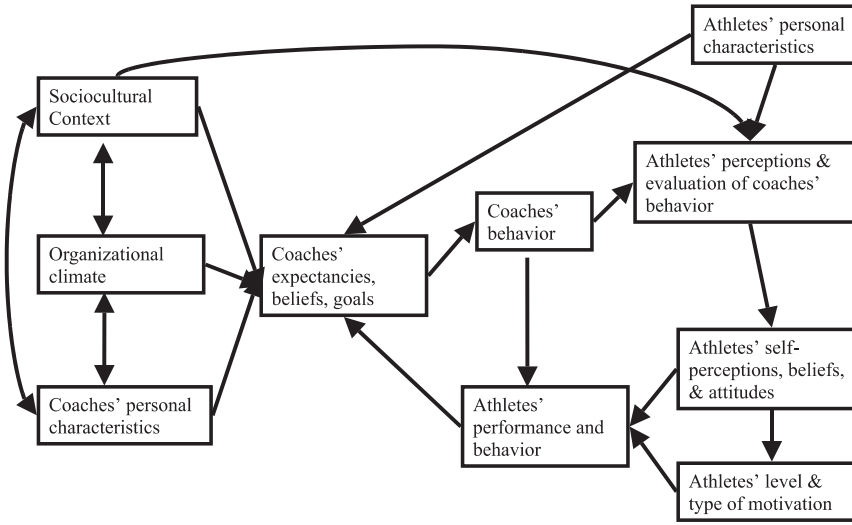

FIGURE 2 Horn's (2002) working model of coaching effectiveness. *Note.* From *Advances in Sport Psychology,* by T. S. Horn (Ed.), 2002 (2nd ed, page 313, figure 10.1), Champaign, IL: Human Kinetics Publishers, Inc. Copyright 2002 by Human Kinetics Publishers, Inc. Reprinted with permission.

INTRODUCTION TO DIFFERENTIAL ITEM FUNCTIONING

The content of this introduction is congruent with the relevant purpose of this article, but it is not an exhaustive treatment of DIF. Readers interested in fuller treatments of DIF are referred to the following resources: Holland and Wainer (1993) for a fairly technical and exhaustive review of DIF in dichotomous data, Camilli and Shepard (1994) for a more applied and targeted treatment of DIF in dichotomous data, and to both Penfield and Lam (2000) and Potenza and Dorans (1995) for a review of DIF in polytomous data.

Over the last few decades, many publications in the education literature have described, applied, and developed procedures for identifying DIF. Because Holland and Wainer (1993) provided a thorough review of much of this work in their volume, the purposes of our introduction are much more modest: to inform readers about what DIF is and to provide a broad overview of the general steps involved in a DIF analysis. In many instances, we attempt to contextualize this introduction in relation to the specific application in this study in an effort to make the information accessible to the intended audience. The given context is not the only context where DIF analysis is possible (e.g., multiple choice items).

A distinction should be made between statistical DIF and substantive DIF (Rousos & Stout, 1996). Statistical DIF occurs when the particular statistical pro-

cedure used flags an item based on an established criteria. In this presentation, the term *DIF* refers to statistical DIF. Substantive DIF occurs after (a) an item has been flagged for DIF, and (b) a construct not entirely relevant to the construct of interest has been determined to be responsible for the observed between-group differences. Substantive DIF will be referred to as bias because items that exhibit substantive DIF can be said to be biased against a particular group. Biased items are a threat to validity because they contribute construct irrelevant variance to the measure of interest (Messick, 1989).

In this study, DIF existed when coaches from different groups (i.e., coach gender or level coached) found a particular CES item (e.g., a ME item) more or less difficult to endorse after conditioning on (i.e., controlling for) the latent efficacy measure (e.g., ME) that was specified to influence responses to the item. For example, suppose that high school coaches were shown to find a specific ME item easier to endorse than college coaches after conditioning on ME (i.e., DIF). Statistically, this meant that on average, of high school and college coaches who had the same level of motivational efficacy, the high school coaches were more likely to assign a higher rating scale category to that item than were the college coaches. Suppose further that it was determined that the reason high school coaches found this item easier to endorse than did college coaches was because of the differing participation motivation of the athletes that high school coaches instruct. In this instance, one could conclude that bias existed because responses to the item appeared to be influenced by characteristics of the athlete, in addition to one's more general belief in his or her ability to affect the psychological mood and psychological skills of athletes. In this scenario, the flagged ME item appears to favor high school coaches over college coaches because of the influence of a nontargeted construct, and therefore, may contribute construct irrelevant variance to ME measures when constructed across these groups.

A thorough and thoughtful DIF analysis, regardless of the type of data or the statistical procedure employed, is unlikely to be driven by a mechanical approach (Clauser & Mazor, 1998). Rather, DIF analysis is a multistep procedure that is dependent on judgment. These judgments should be informed by knowledge of the instrument's development process, relevant theory and research, and practical considerations. General steps common to DIF analyses include (a) selection of groups of examinees to be compared, (b) selection of a criterion on which respondents will be matched, (c) selection of a statistical procedure to identify DIF, (d) computation of DIF indexes, (e) interpretation of the results, and (f) decisions regarding the final makeup of the instrument for particular purposes. These steps are discussed in the paragraphs that follow.

Selecting Groups

Selecting groups of participants to be compared is the first of many judgments to be made by researchers who undertake a DIF analysis. This decision should be supported by theory and or previous research, and should be practical in terms of

group-level sample size. In this study, we selected both coach's gender and level coached as meaningful groupings to be explored. These decisions were consistent with Horn's (2002) model of coaching effectiveness. In any DIF study, as a heuristic (we know of no data-based guidelines for the particular procedure that is to be employed in this study) there should be at least 100 participants per group to achieve reasonably precise estimates. Once substantive and practical groupings are selected, designations of a "reference" and a "focal" group occur. The focal group is the group of primary interest (typically, one against whom there is concern for bias), whereas the reference group is standard to which the focal group is compared. In most cases, including in this study, these designations are arbitrary and only inform the researcher as to how to interpret the direction of observed DIF.

Selecting a Criterion

Selecting a criterion measure on which participants will be matched is a critical decision. If the matching variable is unreliable then the analysis may not be meaningful because responses are being compared between participants who are unevenly matched. That is, we would expect coaches of different efficacy levels to endorse different rating scale categories. Ideally, parallel measures of the same construct (i.e., an external measure) from a different instrument would serve as the criterion measure; however, even in fields where large grants and multiple instruments are sometimes available to researchers, it is relatively uncommon to employ an external criterion measure (Clauser & Mazor, 1998). Rather, an internal criterion measure, such as the measure provided by the instrument under study, is typically employed. Whether internal or external, the criterion measure needs to be a valid measure of the construct of interest. Therefore, evidence of the validity of the measure employed should be provided (as was done earlier), and the reliability of the measures in a particular study should be investigated (as is done later). In this study, the internal criterion measures were subscale scores from the CES which were derived from the multidimensional Rasch rating scale model (to be discussed). Subscale scores, instead of total coaching efficacy scores, were selected because the multidimensional model exhibited better fit to these data than did the unidimensional model (Myers, Wolfe, et al., 2005).

Identify a Differential Item Functioning Procedure

Statistical procedures to identify DIF are so plentiful that a framework for classifying them is necessary (Potenza & Dorans, 1995). Penfield's and Lam's (2000) nontechnical review of this classification system is summarized here. Statistical procedures for identifying DIF can be categorized based on their location on two dimensions: (a) type of ability estimate used as the matching variable, and (b) manner in which item performance at each ability level is determined. Within the first dimension, ability estimates originate from either an observed score methodology,

where the total observed score is the ability estimate, or from a latent variable methodology, where the ability estimate is derived from treating item responses as being indicators of the latent variable of interest. Within the second dimension, estimating item performance at each level of ability incorporates either a parametric methodology, where a mathematical function relates item performance at each level of ability, or a nonparametric methodology, which does not implement a mathematical model to determine observed item performance at each level of observed ability. Dichotomous categorizations within both dimensions yield four classifications of DIF procedures: (a) observed score parametric (e.g., logistic regression), (b) latent variable parametric (e.g., procedures based on IRT models), (c) observed score nonparametric (e.g., Mantel–Haenszel statistic), and (d) latent variable nonparametric (e.g., SIBTEST).

Although approaches from all of the classifications have emerged as valid procedures to identify DIF under varying circumstances, a latent parametric procedure was employed in this study. This choice is warranted because of the several benefits provided by scaling measures within an item response framework. Specifically, IRT (a) allows for routine treatment of missing data, (b) results in measures for both items and respondents that are on a common scale which allows for easier interpretation of measures, (c) specifies several useful diagnostic indexes that can be used to evaluate the performance of individual items and respondents within the measurement framework, and (d) allows for simultaneous scaling of a wide variety of item types, including those that are scored dichotomously, polytomously, and as counts.

Implementing a latent variable parametric statistical procedure for identifying DIF requires investigators to make a number of decisions. First, one must select an appropriate model; we chose a model that fits under the umbrella of IRT. Broadly defined, there are two general traditions of modeling in IRT. First, there are three-parameter and two-parameter models which estimate parameters a (discrimination), b (difficulty), and c (guessing; Hambleton & Swaminathan, 1985; Lord, 1980). Second, there are one-parameter models which estimate parameter b only (Rasch, 1960/1993; Wright & Masters, 1982; Wright & Stone, 1979).[1] In this study, both the three- and two-parameter models were determined to be less appropriate than one-parameter models because estimating the "guessing" parameter for rating scale data makes little sense conceptually, and because the sample size requirement for reasonably precise estimates of a two-parameter model was not met. Lord (1980) offers, as a heuristic, that sample size should equal at least 1,000 when fitting data to a two-parameter IRT model. Therefore one-parameter models, which are also called Rasch (1960/1993) models, were considered.

[1]These classifications are most appropriate for dichotomous data. Classifications of IRT models for polytomous data can be more complicated because of the possibility for additional parameterization (e.g., threshold estimates). Readers who are interested in a fairly applied introduction to this topic are referred to Embretson and Reise (2000).

Smith (2004) presented an overview of DIF detection procedures that are commonly employed within a Rasch measurement framework. In that overview, Smith identified two basic approaches to detecting DIF using the Rasch models. First, Smith referred to the separate calibration *t*-test approach (Lord, 1980; Wright & Masters, 1982; Wright, Mead, & Draba, 1976; Wright & Stone, 1979) which was later referred to as the signed area index by Raju (1988, 1990). In this approach, data from the reference and focal group are calibrated to the Rasch model separately, parameter estimates are rescaled to the same metric, and the item difficulty estimates are compared. Evidence for DIF exists if the item difficulty estimates differ by more than what can be accounted for by estimation error. Second, Smith referred to the "between fit" approach (Smith, 1994; Wright et al., 1976), a procedure in which a between-group item fit statistic is computed. In this approach, a chi-squared fit statistic for each item is created by summing the average of the squared within-group residuals from the Rasch model. Of these approaches, the separate calibration *t* test has been more thoroughly documented and utilized.

In this study, a MRSM (Andrich, 1978) was selected because all of the polytomous items were rated on the same scale and because there was a one-to-one correspondence between category selected and assigned raw score.[2] Conceptually, fitting the data to the MRSM allowed us to estimate the difficulty of each item within each group, put these estimates on the same scale across groups, and compare the difficulty estimates between groups after conditioning on the latent efficacy measure of interest. Finally, implementing any model is often an iterative process if some items exhibit DIF in the initial run. To "purify" the matching criteria, items that exhibited DIF should be removed and the analysis should be rerun.

Computing Differential Item Functioning Indexes

How DIF indexes were computed in this study is a topic that is covered in the next section. Specifically, this topic is covered in the Differential Item Functioning Models and Relative Fit subsections.

---

[2]In instances where polytomous items are not on the same scale or there is not a one-to-one correspondence between category selected and the assigned raw score, a partial credit model may be more appropriate (Masters, 1982). We evaluated the fit of our data to both of these models by comparing the values of the Consistent Akaike Information Criterion (CAIC) for each subscale. For all four of the subscales analyzed in this study, the CAIC indicated that the rating scale model exhibited better fit to these data. Specifically, for the motivation subscale, CAIC = 10034 for the rating scale model and CAIC = 10105 for the partial credit model; for the GS subscale, CAIC = 10151 for the rating scale model and CAIC = 10213 for the partial credit model; for the technique subscale, CAIC = 8717 for the rating scale model and CAIC = 8720 for the partial credit model; for the character building subscale, CAIC = 5218 for the rating scale model and CAIC = 5242 for the partial credit model.

Interpreting the Results

Interpreting the results and making final decisions about the make-up of the instrument for a particular purpose are intertwined processes. Both of these processes need to be influenced by theory and previous research. Initial steps in the interpretation process should include (a) investigating the content and history of the flagged item, and (b) considering any substantive differences between the groups that may explain the observed DIF. Investigating the content of the item should include comparing it to other items intended to measure the construct and comparing its fit with the operational definition of the construct. Investigating the history of the item should include determining how it has operated in similar applications, if available. Both investigations should be directed toward determining if a case can be made that the identified item measures another ability in addition to the ability of interest. Considering substantive differences between the groups that may explain the DIF highlights the range of possible interpretations for explaining why an item exhibited DIF. For instance, Sheuneman and Subhiyah (1998) explored DIF on a certification exam for a medical specialty based on qualification pathway (i.e., residency vs. equivalency) and by specialty area (i.e., specialty I vs. specialty II) groupings. After investigating the content of the items and considering substantive differences between the groups, these researchers argued that the items that exhibited DIF did so because of known differences between training experiences. Therefore, they argued that the identified items did not represent a validity threat (i.e., bias); rather, the identified items represented validity evidence of the substantive differences in training experiences between groupings.

Decision Making

Final decisions about the make-up of the instrument should be driven by the intended purpose of the instrument. In addition, these final decisions should be congruent with the interpretations in the previous step and be practical. Decisions about items that exhibit DIF can include the following: retain the item for substantive reasons as in the Sheuneman and Subhiyah (1998) example, drop the biased item in the current format and suggest specific revisions based on the interpretation of why the item exhibited DIF, and drop the biased item because it measures an additional ability that is irrelevant to the ability of interest. Whatever the recommendation, a rationale must be provided. Last, decisions about the final make-up of the instrument should be practical. For instance, items that exhibit DIF based on a given grouping bring into question the validity of measures derived from samples that included participants from both of these groups. One practical way to address this concern, while retaining all of the items, is to construct measures separately for each group; however, invoking this strategy precludes comparing estimates between groups, as measures would not have the same origin. If comparing estimates

between identified groups is important to the purpose of a study, then dropping the items that exhibited DIF based on the grouping variable of interest is probably appropriate. Because dropping items may reduce the content validity of the instrument, researchers may consider targeting only one of the groups within a study to minimize validity concerns for the measures that result.

MRSM AND AN ASSOCIATED PROCEDURE FOR IDENTIFYING DIFFERENTIAL ITEM FUNCTIONING OF POLYTOMOUS DATA

This introduction consists of three parts. First, the basic components of the MRSM are briefly described, ignoring the specifics of a DIF analysis (see Tenenbaum & Fogarty, 1998, for a fuller introduction to the Rasch model within an exercise science context). Second, the models employed to test for DIF in this study are defined. Third, how we evaluated the relative fit of competing models is made clear.

Basic Components of MRSM

In this study, four unidimensional measurement scales were created; one for each dimension of the CES. To construct measures for each dimension of the CES, we employed a variation of the MRSM which described the probability that a specific coach (*n*) would rate a particular item (*i*) using a specific rating scale category (*k*), conditioned on the coach's efficacy of interest (θ_n) and the item's difficulty (δ_i). Thus, the log-odds equation for this probability,

$$\log(P_{nik} / P_{nik-1}) = \theta_n - \delta_i - \tau_k, \quad (1)$$

contained three parameters: θ_n, δ_i, and category threshold (τ_k), the difficulty of surpassing the threshold between two adjacent rating scale categories. Note that the parameters estimated in this model represented characteristics of the coach, item, and rating scale, and that a linear combination of these parameters defined the log-odds (logit—the natural logarithm of the odds of an event) of a particular coach with a given level of efficacy assigning a rating in a particular category versus assigning a rating in the next lower category to an item with a given difficulty. Parameters were estimated via a Monte Carlo implementation of the expected-maximization algorithm (Wu, Adams, & Wilson, 1998). For identification purposes, mean δs were constrained to equal zero. Thus, negative δ estimates indicated that an item was relatively easy to endorse, whereas positive δ estimates indicated that an item was relatively difficult to endorse.

Differential Item Functioning Models

DIF analyses were performed to determine whether the CES items exhibited DIF based on coach gender or level coached on each of the four CES dimensions using two nested models. The most complex model was as follows:

$$\log(P_{nigk}/P_{nigk-1}) = \theta_n - \delta_i - \mu_g - \tau_k - \beta_{ig}, \quad (2)$$

where

θ_n was the coach's efficacy of interest,
δ_i was the item difficulty (i.e., item main effect),
μ_g was the mean calibration for each group (i.e., group main effect),
τ_k was the rating scale category threshold, and
β_{ig} was the item-by-group interaction.

The inclusion of μ_g in the model removed the influence of group differences in coaching efficacy from the interpretation of interactions between groups and item difficulty, β_{ig}.

It is instructive to define what the interaction term, β_{ig}, represented. This term represented the deviation of the item difficulty for each group from the overall item difficulty, δ_i. Hence, β_{ig} was analogous to what Smith (2004) referred to as the "separate calibration" approach to DIF detection with the Rasch model. In fact, the value of β_{ig} directly corresponded to the relevant effect size index within that framework. Specifically, Raju's Signed Area Index (SAI) is computed as the difference between separate calibrations of item difficulty (δ_i) using data from each group (i.e., SAI = $\delta_{i\textit{reference}} - \delta_{i\textit{focal}}$, whereas $\beta_{ig} = \delta_i - \delta_{i|g}$, so that β_{ig} = SAI / 2). Hence, if the absolute value of β_{ig} were large, there was evidence of DIF for item *i* because the difficulty of the item differed for the two groups after controlling for differences in coaching efficacy. This model is referred to as Model 1 from this point forward.[3]

To evaluate the contribution of β_{ig}, Model 1 was compared to a model which assumed that the items functioned similarly for both groups (i.e., Model 2):

$$\log(P_{nigk}/P_{nigk-1}) = \theta_n - \delta_i - \mu_g - \tau_k, \quad (3)$$

Comparing these two models provided an omnibus test for DIF across items within a subscale. Because this was an omnibus test, post hoc analyses were performed to

[3]Note that this model assumes that the rating scale itself functions similarly for both groups (i.e., there is no differential rating scale functioning). Differential rating scale functioning is an area worthy of future research but is beyond the purview of this article.

identify item-level DIF by examining the effect size between item difficulties for the two groups via the SAI. Items were flagged if the absolute difference between an item's difficulties for the two groups was ≥ .50 logits (Scheuneman & Subhiyah, 1998).[4]

Relative Fit

Relative fit of the nested models was evaluated using a likelihood ratio chi-square statistic (χ^2_{LR}). The χ^2_{LR} statistic was used to determine whether the difference between the fit of the two models was greater than could be attributed to estimation error (Thissen, Steinberg, & Wainer, 1993),

$$\chi^2_{LR}\ (df) = [-2LN(Likelihood_{complex})] - [-2LN\ (Likelihood_{simple})], \qquad (4)$$

where

–2LN (Likelihood) was the deviance index for the model in question, and
df was the degrees of freedom for the likelihood ratio chi-squared test, equal to the
difference of the number of parameters in the two models.

Because the χ^2_{LR} test is sensitive to sample size, CAIC (Bozdogan, 1987) values were also considered:

$$CAIC = [\ -2LN(Likelihood)] + p\ [1+ LN(n)] \qquad (5)$$

where

n is the sample size, and
p is the number of parameters estimated for the model.

The CAIC depicted the fit of the model in question relative to the number of parameters estimated, where lower values indicated better fit (Wicherts & Dolan, 2004).

---

[4]It is possible to perform a hypothesis test on each item by dividing the item's item-by-group interaction parameter estimate by its standard error to obtain a Wald *t* test statistic. We did not utilize this particular statistic because these hypothesis test statistics tend to be sensitive to sample size. Rather, we performed an omnibus hypothesis test and then examined an effect size indicator for each item as a follow-up to that test. Our approach identified fewer items as exhibiting DIF than would have been flagged based on hypothesis test statistics. Standard errors for the values of β_{ig}, however, are reported in this article (see Tables 3 and 4).

AN APPLICATION

The remainder of this article is concerned with purpose (c): applying this procedure to previously collected data from American coaches who responded to the CES. Specifically, the following research questions were examined:

1. Did the CES items exhibit DIF based on coach gender?
2. Did the CES items exhibit DIF based on level coached?

METHOD

Sample

Data for this study were collated from all of the published studies with sample sizes greater than 50 that have employed the CES with American coaches (Feltz et al., 1999; Malete & Feltz 2000; Myers, Vargas-Tonsing, et al., 2005). When demographics were available, cases were coded based on level coached, coach's gender, and coach's race. These data are illustrated in Table 1. Sport coached and special concerns within datasets were also noted and are summarized later. Datasets were then combined across studies ($N = 665$).

The Feltz et al. (1999) data provided two independent samples. Sample 1 coaches represented the sports of basketball (29%), track (13%), volleyball (11%), cross-country running (7%), baseball (7%), tennis (7%), and other sports.[5] Sample 2 participants coached basketball (26%), volleyball (13%), track (11%), football (11%), softball (6%), and other sports. The Malete and Feltz (2000) data were collected within a coaching education program and included premeasures and postmeasures. Only preprogram data were retained in this study to avoid problems with dependency. Participants coached basketball (18%), football (12%), cheerleading (12%), soccer (7%), softball (7%), baseball (7%), and other sports. The Myers, Vargas-Tonsing, et al. (2005) data consisted of participants who coached softball (26%), baseball (20%), soccer (34%), and basketball (21%) at non-Division I collegiate levels.

The Myers, Wolfe, et al. (2005) study and this study summarized the results of a single large project. The previous study provided validity evidence for the criterion measures used in this study (i.e., evidence for the fit of the internal model and evidence for item-level fit) and will not be reprinted here. One empirically based decision implemented in the previous study directly affected this study, however, and is briefly reviewed here. In the previous study, rating scale analyses suggested that coaches were asked to distinguish between too many levels of coaching efficacy in

[5]Each of the "other" sports comprised ≤ 5% of the participants.

TABLE 1
Demographic Information Within and Across Studies

	Feltz, Chase, Moritz, and Sullivan (1999)				*Malete and Feltz (2000)*		*Myers, Vargas-Tonsing, and Feltz (2005)*		*Total Sample*	
	N = 188		*N = 291*[a]		*N = 60*		*N = 126*		*N = 665*	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Level coached					NA = 24	40			NA = 24	4
Youth					21	35			21	3
High school	188	100	291	100	15	25			494	74
Collegiate							126	100	126	19
Gender										
Male	109	58	163	56	34	57	84	67	227	61[b]
Female	79	42	128	44	26	43	42	33	147	39[b]
Race	NA = 9	5	NA = 47	16	NA = 4	6	NA = 12	10	72	11
African American					10	17	4	3	14	2
White	179	95	244	84	46	77	110	87	579	87

Note. NA = not available.

[a]Complete demographic data were not provided in the dataset forwarded. Information listed is based on statistics reported in the relevant manuscript. [b]Demographic data are based on demographic data attributable to specific cases ($N = 374$).

the original rating scale structure (10 categories), and that the data should be, and were, collapsed into four categories. The same data structure was maintained in this study. In both studies, the four category structure was interpreted as indicating low, moderate, high, and complete confidence.

Analyses

Internal consistency reliability. The consistency of rank orderings of coaching efficacy estimates across measurement contexts was examined with reliability of separation coefficients (α). The reliability of separation coefficient is analogous to Cronbach's (1951) alpha, but it is based on estimates of true and error variance derived from IRT models. Specifically, the reliability of separation for efficacy estimates is equal to

$$[V(\hat{\theta}) - MSE(\hat{\theta})] / V(\hat{\theta}), \tag{6}$$

where

$V(\hat{\theta})$ is the variance of the efficacy estimates, and
$MSE(\hat{\theta})$ is the mean error variance of the efficacy estimates.

This equation is comparable to the true score test theory definition of reliability as the ratio of true variance to observed variance. As suggested by Kline (1998) and in relation to the purpose of this study, αs greater than .90 were considered excellent, αs greater than .80 were considered very good, and αs greater than .70 were considered adequate.

Differential item functioning. DIF analyses were implemented and interpreted as detailed earlier. Alpha was set equal to .05 for all hypothesis tests. In a few cases, the results of the omnibus χ^2_{LR} test and the CAIC values conflicted; that is, a statistically significant effect was not meaningfully large according to CAIC values. Congruent with the purposes of this article, in all cases, item-level DIF was explored for the sake of illustration. It is noted that a nonsignificant omnibus test for DIF could be used as justification for forgoing an exploration of item-level DIF. Items flagged for exhibiting DIF were dropped in a subsequent DIF analysis to "purify" the criterion measure. Hypotheses for why flagged items exhibited DIF, and final recommendations regarding future use of those items, are provided later.

For the gender analysis, women ($n = 149$) were specified as the reference group and men ($n = 224$) as the focal group.[6] For the level analysis, high school coaches ($n = 539$) were specified as the reference group and collegiate coaches ($n = 126$) as

[6]Complete demographic data were not provided in the Feltz et al. (1999) dataset.

the focal group. In accordance with the internal model, analyses were applied to responses to ME, GS, TE, and CB items, separately.

RESULTS

Descriptive Statistics

Table 2 depicts descriptive statistics of the logit-based measures. Logit-based, opposed to raw scores, were selected based on the purposes of this study. In all cases, the distributions of measures approximated normal distributions. In addition, the reliability coefficients (.91 for ME, .91 for GS, .90 for TE, and .83 for CB) suggested good levels of internal consistency and afforded confidence in the validity of the matching variables.

Coach Gender

Motivation Efficacy (ME). The omnibus test for DIF, β_{ig}, was statistically significant $\chi^2_{LR}(6, N = 374) = 17.97$, $p = .006$, but the CAIC values indicated that Model 2 (CAIC = 5657.61) fit the data at least as well as Model 1 (CAIC = 5681.18). Only one of the ME item-level difficulties varied meaningfully based on coach's gender: ME3, "mentally prepare athletes for game/meet strategies," *SAI* = .63. Women found ME3 more difficult to endorse than did men, after conditioning on ME.

A subsequent DIF analysis was run on the six ME items that did not exhibit DIF to purify the measure. The omnibus test for DIF, β_{ig}, was not statistically significant $\chi^2_{LR}(5, N = 374) = 4.41$, $p = .492$, and the CAIC values indicated that Model 2 (CAIC

TABLE 2
Descriptive Statistics of Logit-Based Measures

	Latent Factor			
	Motivation Efficacy	*Game Strategy Efficacy*	*Technique Efficacy*	*Character Building Efficacy*
M	–0.08	0.26	0.75	1.78
SD	1.84	1.87	1.80	2.10
Range	–5.24 to 4.54	–5.49 to 4.87	–4.77 to 5.16	–4.86 to 5.91
Skewness (*SE*)	.17 (.10)	.07 (.10)	.08 (.10)	–.14 (.10)
Kurtosis (*SE*)	–.11 (.19)	–.05 (.19)	.06 (.19)	–.38 (.19)
Reliability	0.91	0.91	0.90	0.83
Mean *SE*	0.31	0.30	0.33	0.73

TABLE 3
Differential Item Functioning by Coach Gender

Subscale	β_{ig}	*SE*	*SAI*
Motivation Efficacy (ME)			
ME 1	–0.01	0.07	–0.03
ME 3	—	—	—
ME 6	–0.09	0.07	–0.17
ME10	0.12	0.07	0.24
ME12	–0.11	0.07	–0.23
ME15	0.11	0.07	0.21
ME 23	–0.01	—	–0.02
Game Strategy Efficacy (GS)			
GS2	0.12	0.06	0.25
GS4	0.03	0.06	0.06
GS8	–0.16	0.06	–0.31
GS9	–0.04	0.06	–0.09
GS11	0.03	0.06	0.07
GS17	—	—	—
GS21	0.02	—	0.03
Technique Efficacy (TE)			
TE7	–0.02	0.06	–0.04
TE14	0.01	0.06	0.02
TE16	–0.11	0.06	–0.22
TE18	0.15	0.07	0.30
TE20	0.07	0.06	0.15
TE22	0.02	—	0.04
Character Building Efficacy (CB)			
CB5	–0.04	0.07	–0.09
CB13	0.07	0.07	0.14
CB19	–0.07	0.08	–0.13
CB24	0.04	—	0.08

Note. β_{ig} = item by group interaction term; SAI = Signed Area Index.

= 4808.54) fit the data at least as well as Model 1 (CAIC = 4838.75). None of the ME item-level difficulties exhibited DIF based on coach's gender (see Table 3).[7]

Game Strategy Efficacy (GS). The omnibus test for DIF, β_{ig}, was statistically significant $\chi^2_{LR}(6, N = 374) = 12.74$, $p = .047$, but the CAIC values indicated that Model 2 (CAIC = 5687.09) fit the data at least as well as Model 1 (CAIC = 5715.90). Only one of the GS item-level difficulties varied meaningfully based on coach's gender: GS17, "maximize your team's strengths during competition," *SAI*

[7]Tables 3 and 4 summarize information from models where DIF did not exist. In some cases this meant that one or two items were dropped to purify the measures. Results from the initial models where DIF was observed are available from the lead author on request.

= –.51. Men found GS17 more difficult to endorse than did women, after conditioning on GS.

A subsequent DIF analysis was run on the six GS items that did not exhibit DIF to purify the measure. The omnibus test for DIF, β_{ig}, was not statistically significant $\chi^2_{LR}(5, N = 374) = 4.64, p = .461$, and the CAIC values indicated that Model 2 (CAIC = 4986.14) fit the data at least as well as Model 1 (CAIC = 5016.13). None of the ME item-level difficulties exhibited DIF based on coach's gender (see Table 3).

Technique Efficacy (TE). The omnibus test for DIF, β_{ig}, was not statistically significant $\chi^2_{LR}(5, N = 374) = 3.49, p = .625$, and the CAIC values indicated that Model 2 (CAIC = 4915.37) fit the data at least as well as Model 1 (CAIC = 4946.50). None of the TE item-level difficulties varied meaningfully based on coach's gender (see Table 3).

Character Building Efficacy (CB). The omnibus test for DIF, β_{ig}, was not statistically significant $\chi^2_{LR}(3, N = 374) = 0.93, p = .818$, and the CAIC values indicated that Model 2 (CAIC = 3028.09) fit the data at least as well as Model 1 (CAIC = 3047.94). None of the CB item-level difficulties varied meaningfully based on coach's gender (see Table 3).

Level Coached

Motivation Efficacy (ME). The omnibus test for DIF, β_{ig}, was statistically significant $\chi^2_{LR}(6, N = 665) = 42.52, p < .001$, but the CAIC values indicated that Model 2 (CAIC = 10009.95) fit the data as well as Model 1 (CAIC = 10012.43). Two of the ME item-level difficulties varied meaningfully based on level coached: ME3, "mentally prepare athletes for game/meet strategies," SAI = .60; and ME12, "build team cohesion," SAI = .59. High school coaches found ME3 and ME12 more difficult to endorse than did college coaches, after conditioning on ME.

A subsequent DIF analysis was run on the five ME items that did not exhibit DIF to purify the measure. The omnibus test for DIF, β_{ig}, was not statistically significant $\chi^2_{LR}(4, N = 665) = 4.89, p = .299$, and the CAIC values indicated that Model 2 (CAIC = 7203.65) fit the data at least as well as Model 1 (CAIC = 7226.47). None of the ME item-level difficulties exhibited DIF based on level coached (see Table 4).

Game Strategy Efficacy (GS). The omnibus test for DIF, β_{ig}, was not statistically significant $\chi^2_{LR}(6, N = 665) = 4.15, p = .656$, and the CAIC values indicated that Model 2 (CAIC = 10083.82) fit the data at least as well as Model 1 (CAIC = 10196.17). None of the GS item-level difficulties varied meaningfully based on level coached (see Table 4).

Technique Efficacy (TE). The omnibus test for DIF, β_{ig}, was not statistically significant χ^2_{LR} (5, N = 665) = 2.08, p = .838, and the CAIC values indicated that Model 2 (CAIC = 8724.33) fit the data as well as Model 1 (CAIC = 8759.75). None of the TE item-level difficulties varied meaningfully based on level coached (see Table 4).

Character Building Efficacy (CB). The omnibus test for DIF, β_{ig}, was not statistically significant χ^2_{LR} (3, N = 665) = 6.81, p = .078, and the CAIC values indicated that Model 2 (CAIC = 5220.69) fit the data as well as Model 1 (CAIC = 5236.38). None of the CB item-level difficulties varied meaningfully based on level coached (see Table 4).

TABLE 4
Differential Item Functioning by Level Coached

Subscale	β_{ig}	*SE*	*SAI*
Motivation Efficacy (ME)			
ME 1	–0.09	0.08	–0.19
ME 3	—	—	—
ME 6	–0.06	0.07	–0.13
ME10	–0.04	0.08	–0.09
ME12	—	—	—
ME15	0.11	0.07	0.22
ME 23	0.09	—	0.18
Game Strategy Efficacy (GS)			
GS2	–0.04	0.07	–0.08
GS4	0.08	0.07	0.15
GS8	0.00	0.07	–0.01
GS9	0.05	0.07	0.10
GS11	0.07	0.07	0.15
GS17	–0.08	0.07	–0.15
GS21	–0.08	—	–0.16
Technique Efficacy (TE)			
TE7	0.01	0.07	0.02
TE14	0.02	0.07	0.04
TE16	0.06	0.07	0.12
TE18	–0.07	0.07	–0.14
TE20	–0.04	0.07	–0.09
TE22	0.02	—	0.04
Character Building Efficacy (CB)			
CB5	–0.07	0.08	–0.13
CB13	0.13	0.08	0.25
CB19	–0.16	0.08	–0.33
CB24	0.11	—	0.21

Note. β_{ig} = item by group interaction term; SAI = Signed Area Index.

DISCUSSION

This study (a) provides a conceptual introduction to DIF, (b) introduces the MRSM and an associated statistical procedure for identifying DIF in rating scale items, and (c) applies this procedure to previously collected data from American coaches who responded to the CES. The first two purposes are accomplished in previous sections and are not discussed further. Interpreting the results of the final purpose, which is concerned with the generalizability of CES measures across substantive subgroups of coaches, is the focus of this section.

A CES item displays DIF if coaches from different groups are more or less likely to endorse that item once coaches are matched on the efficacy of interest, where ME, GS, TE, and CB efficacies define coaching efficacy. Coach gender and level coached are selected as the grouping variables. The vast majority of items did not exhibit DIF based on coach gender (92%) or level coached (92%). None of the TE and CB items exhibit DIF based on coach gender or level coached. One of the ME items and one of the GS items exhibit DIF based on coach gender. Two of the ME items exhibit DIF based on level coached.

That the vast majority of items did not exhibit DIF based on coach gender provides evidence for much of the internal model depicted in Figure 1. That is, responses to the said items are specified to be uninfluenced by coach gender after controlling for the relevant ability level, and evidence is provided to support that assumption for the vast majority of items. Practically, this finding suggests that these measures, less the two items that did exhibit DIF, can be constructed and compared (i.e., they generalize) across both groups. This assumption has been embedded, without empirical justification, in previous literature (Feltz et al., 1999; Lee et al., 2002; Malete & Feltz, 2000; Myers, Vargas-Tonsing, et al., 2005; Myers, Wolfe, et al., 2005; Sullivan & Kent, 2003; Vargas-Tonsing et al., 2003). Future users of the CES now have some empirical support for assuming that these measures, less the two items that did exhibit DIF, can be constructed and compared across both genders because it appears that the items maintain their relative difficulties across both groups.

There is evidence that one of the ME items, ME3—"mentally prepare athletes for game/meet strategies," exhibits DIF based on coach gender. Women find ME3 more difficult to endorse than did men, after conditioning on ME. Speculations as to why female coaches find this item more difficult to endorse are guided by literature on perceived coaching competence (Weiss, Barber, Sisley, & Ebbeck, 1991). Weiss et al. (1991) reported that novice female coaches identified a negative aspect of coaching to be an overemphasis on winning, and their own leadership skills as a weakness. These same coaches identified interpersonal communication and social support as primary strengths. It may be that female coaches view the relationship-based aspects of coaching as a central responsibility and spend more time developing these skills and confidences as opposed to skills focused solely on men-

tally preparing athletes for competition. Another explanation is that female coaches could find mentally preparing athletes for competition difficult to endorse because of inexperience in such situations due to attrition of female coaches (Acosta & Carpenter, 2000; Weiss & Stevens, 1993). It may be that experience in mentally preparing athletes for competitive situations, and not gender, explains why male coaches found this item easier to endorse than did female coaches, after conditioning on ME. Although the degree to which these speculations are tenable can be debated, we view ME3 as biased in this scenario and recommend excluding it when wishing to construct and compare unbiased measures of ME across coach gender. Because there is also evidence for the utility of the item as an indicator of ME (Feltz et al., 1999; Lee et al., 2002; Myers, Wolfe, et al., 2005), we recommend including the item when not making direct comparisons between these two groups.

There is evidence that one of the GS items, GS17—"maximize your team's strengths during competition," exhibits DIF based on coach gender. Men find GS17 more difficult to endorse than did women, after conditioning on GS. Given the speculations in the previous paragraph, which were based on the relevant research, why male coaches find this item more difficult to endorse is not clear. The relevant previous research would predict the opposite of what was observed, if any difference was to be expected. Given the lack of explanation as to why DIF was observed, coupled with the fact that this item barely met our cutoff for DIF, we suggest that this finding may have been spurious and should be viewed with caution. Despite these suggestions, until contrary evidence is available, we view GS17 as inappropriate in this scenario and recommend excluding it when wishing to construct and compare measures of GS across coach gender. Because there is also evidence for the utility of the item as an indicator of GS (Feltz et al., 1999; Lee et al., 2002; Myers, Wolfe, et al., 2005), we recommend including the item when not making direct comparisons between these two groups.

That the vast majority of items did not exhibit DIF based on level coached provides evidence for much of the internal model depicted in Figure 1. That is, responses to the said items are specified to be uninfluenced by level coached after controlling for the relevant ability level, and evidence is provided to support that assumption for the vast majority of items. Practically, this finding suggests that these measures, less the two items that did exhibit DIF, may generalize across coaches of high school and lower division collegiate athletes. This assumption has been embedded, without empirical justification, in previous literature (Myers, Wolfe, et al., 2005). Future users of the CES now have some empirical support for assuming that these measures, less the two items that did exhibit DIF, can be constructed and compared across these groups because it appears that the items maintain their relative difficulties across both groups.

There is evidence that two of the ME items exhibit DIF based on level coached. High school coaches find ME3, <@147>mentally prepare athletes for game/meet strategies," and ME12, "build team cohesion," more difficult to endorse than did

collegiate coaches, after conditioning on ME. Speculation as to why high school coaches find mental preparation of athletes and building team cohesion more difficult to endorse is guided by consideration of the developmental stage and possible motivations for participation of high school athletes. It may be that high school coaches find their athletes' psychological skills and their interest in a cohesive team to be quite variable and difficult to influence, and that this variability and resistance to outside influence are partially due to why high school athletes participate in sport (e.g., for fun). We view both of the identified items as biased in this scenario and recommend dropping them when wishing to construct and compare unbiased measures of ME across high school coaches and college coaches. Because there is also evidence for the utility of both items as indicators of ME (Feltz et al., 1999; Lee et al., 2002; Myers, Wolfe,et al., 2005), we recommend including both items when not making direct comparisons between these two groups.

Three limitations of the data used in this study are noted. First, the statistical power of both sets of analyses was moderate. Future research with larger sample sizes would provide more sensitive investigation. Second, sport coached varied across groups. Future research that explores the possibility of DIF based on sport coached would extend validity evidence for the CES. This point is especially salient because in the vast majority of relevant studies, a diversity of sports was represented, but it was not modeled when creating efficacy measures (Feltz et al., 1999; Lee et al., 2002; Malete & Feltz, 2000; Myers, Vargas-Tonsing, et al., 2005; Sullivan & Kent, 2003). Clearly the assumption has been that sport coached does not affect how coaches respond to the CES items, after controlling for the efficacy of interest. Future researchers should empirically test this assumption. Third, the matching of coach gender and athlete gender varied. That is, female coaches coached female athletes but male coaches coached male and female athletes. There is evidence that in female teams, coaching efficacy may operate differently for male and female coaches (Myers, Vargas-Tonsing, et al., 2005). Future researchers should investigate the degree to which the CES items may exhibit DIF based on the matching of coach and athlete gender.

Given the burgeoning role of coaching efficacy in coaching education research, that the CES is the only instrument purported to measure the construct, and the validity evidences forwarded in this study, continued use of the CES for the intended purposes appears to be reasonable. The intended purposes include obtaining measures to determine sources of coaching efficacy, to examine the influence of coaching efficacy on athlete and team variables, and to assess the ability of education programs to alter coaching efficacy. The evidence provided in this study largely supports the use of the CES for making direct comparisons of substantively interesting subgroups of coaches (i.e., male and female coaches and coaches of high school and lower level college athletes) on the dimensions of coaching efficacy. Of course, users of the CES should take into account the validity concerns highlighted in this manuscript when imposing measurement models across relevant subgroups

of coaches. Further research is needed to determine how the items function for coaches of youth sports.

ACKNOWLEDGMENTS

This study was supported in part by a William Wohlgamuth Memorial Scholarship for the Study of Youth in Sports at Michigan State University. We gratefully acknowledge Ahnalee Brincks for editing this manuscript.

REFERENCES

Acosta, R. V., & Carpenter, L. J. (2000). *Women in intercollegiate sport: A longitudinal study — twenty-three year update, 1977–2000.* Brooklyn, NY: Brooklyn College, Department of Physical Education.

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–41.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Horn, T. S. (2002). Coaching effectiveness in the sports domain. In T. S. Horn (Ed.), *Advances in sport psychology* (pp. 309–354). Champaign, IL: Human Kinetics.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Lee, K. S., Malete, L., & Feltz, D. L. (2002). The effect of a coaching education program on coaching efficacy. *International Journal of Applied Sport Sciences, 14,* 55–67.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Malete, L., & Feltz, D. L., (2000). The effect of a coaching education program on coaching efficacy. *The Sport Psychologist, 14,* 410–417.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Muraki, E. (1993, April). *Implementing item parameter drift and bias in polytomous item response models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Myers, N. D., Vargas-Tonsing, T. M., & Feltz, D. L. (2005). Coaching efficacy in collegiate coaches: Sources, coaching behavior, and team variables. *Psychology of Sport & Exercise, 6,* 129–143.

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.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5–15.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23–37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press. (Reprinted from *Probabilistic models for some intelligence and attainment tests,* by G. Rasch, 1960, Copenhagen: Danish Institute for Educational Research)

Rogers, H. J., & Swaminathan, H. (1993, April). *Logistic regression procedures for detecting DIF in non-dichotomous item responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Rousos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Scheuneman, J. D., & Subhiyah, R. G. (1998). Evidence for the validity of a Rasch model technique for identifying differential item functioning. *Journal of Outcome Measurement, 2,* 33–42.

Smith, R. M. (1994). A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement, 54,* 886–896.

Smith, R. M. (2004). Detecting item bias with the Rasch model. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 391–418). Maple Grove, MN: JAM Press.

Sullivan, P. J., & Kent, A. (2003). Coaching efficacy as a predictor of leadership style in intercollegiate athletics. *Journal of Applied Sport Psychology, 15,* 1–11.

Tenenbaum, G., & Fogarty, G. (1998). Application of the Rasch analysis to sport and exercise psychology measurement. In J. L. Duda (Ed.), *Advancements in sport and exercise psychology measurement* (pp. 409–421). Morgantown, WV: Fitness Information Technology.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vargas-Tonsing, T. M., Warners, A. L., & Feltz, D. L. (2003). The predictability of coaching efficacy on team efficacy and player efficacy in volleyball. *Journal of Sport Behavior, 26,* 396–407.

Weiss, M. R., Barber, H., Sisley, B. L., & Ebbeck, V. (1991). Developing competence and confidence in novice female coaches: II. Perceptions of ability and affective experiences following a season-long coaching internship. *Journal of Sport & Exercise Psychology, 13,* 336–363.

Weiss, M. R., & Stevens, C. (1993). Motivation and attrition of female coaches: An application of social exchange theory. *The Sport Psychologist, 7,* 244–261.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1–19.

Wicherts, J. M., & Dolan, C. V. (2004). A cautionary note on the use of information fit indices in covariance structure modeling with means. *Structural Equation Modeling, 11,* 45–50.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B. D., Mead, R. J., & Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model* (Research Memorandum 22). Chicago: University of Chicago, Department of Education, MESA Statistical Laboratory.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ACER ConQuest: Generalized Item Response Modeling Software (Version 1.0) [Computer software]. Melbourne, Victoria, Australia: Australian Council for Educational Research.

APPENDIX

How confident are you in your ability to—-

1. maintain confidence in your athletes? (ME1)
2. recognize opposing team's strengths during competition? (GS2)
3. mentally prepare athletes for game/meet strategies? (ME3)
4. understand competitive strategies? (GS4)
5. instill an attitude of good moral character? (CB5)
6. build the self-esteem of your athletes? (ME6)
7. demonstrate the skills of your sport? (TE7)
8. adapt to different game/meet situations? (GS8)
9. recognize opposing team's weakness during competition? (GS9)
10. motivate your athletes? (ME10)
11. make critical decisions during competition? (GS11)
12. build team cohesion? (ME12)
13. instill an attitude of fair play among your athletes? (CB13)
14. coach individual athletes on technique? (TE14)
15. build the self-confidence of your athletes? (ME15)
16. develop athletes' abilities? (TE16)
17. maximize your team's strengths during competition? (GS17)
18. recognize talent in athletes? (TE18)
19. promote good sportsmanship? (CB19)
20. detect skill errors? (TE20)
21. adjust your game/meet strategy to fit your team's talent? (GS21)
22. each the skills of your sport? (TE22)
23. build team confidence? (ME23)
24. instill an attitude of respect for others? (CB24)