ELSEVIER

Psychology of Sport and Exercise 2 (2001) 1–26

Psychology
OF SPORT AND EXERCISE

www.elsevier.com/locate/psychsport

Analysis of the affect measurement conundrum in exercise psychology: II. A conceptual and methodological critique of the Exercise-induced Feeling inventory

Panteleimon Ekkekakis [a,*], Steven J. Petruzzello [b]

[a] *253 Forker Building, Department of Health and Human Performance, Iowa State University, Ames, IA 50011, USA*

[b] *Department of Kinesiology, University of Illinois at Urbana-Champaign, USA*

Received 7 October 1999; received in revised form 28 August 2000; accepted 11 September 2000

Abstract

Background and purpose: The measurement of affect in the context of exercise has become a controversial issue. To help elucidate some of the problems, the conceptual and methodological bases of the Exercise-induced Feeling Inventory (Gauvin, L., & Rejeski, W.J. (1993). The Exercise-induced Feeling Inventory: Development and initial validation. *Journal of Sport & Exercise Psychology, 15,* 403.) are critiqued, emphasizing deviations from established scale development and validation guidelines.

Methods: From a conceptual standpoint, the analysis concentrates on the definition of "feeling states," the demarcation of the content domain, the decision to adopt a categorical conceptualization of affect, the notion of a universal phenomenology of exercise, and the notion of exercise-specific affect. From a methodological standpoint, emphasis is placed on the item selection and content validation methods, the exploratory structural analysis, and the application of structural equation modeling.

Results and conclusions: Substantial deficiencies in conceptual groundwork and deviations from established guidelines are identified that may have important implications for the validity and utility of the EFI.

Keywords: Feeling states; Structure of affect; Phenomenology; Scale development

In a companion paper (Ekkekakis & Petruzzello, 2000), we examined a number of important issues surrounding the measurement of affect in the context of exercise and identified several reasons for concern. Because measurement is the foundation of any research endeavor, the close

* Corresponding author. Tel.: +1-515-294-8766; fax: +1-515-294-8740.
E-mail address: ekkekaki@iastate.edu (P. Ekkekakis).

1469-0292/01/$ - see front matter
PII: S1469-0292(00)00020-0

inspection and resolution of these problems is necessary before substantial progress can be made in the study of the exercise-affect relationship.

The present paper extends our analysis by focusing on a specific measure, namely the Exercise-induced Feeling Inventory (EFI; Gauvin & Rejeski, 1993). The EFI was the product of escalating dissatisfaction with traditionally used measures of affective variables in the context of exercise, such as the Profile of Mood States (POMS; McNair, Lorr & Droppleman, 1971) and the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch & Lushene, 1970). The EFI is a 12-item measure of "exercise-induced feeling." It includes four subscales, namely Positive Engagement, Revitalization, Tranquility, and Physical Exhaustion. The popularity of the scale is growing as evidenced by its adoption by independent researchers (e.g., Annesi & Mazas, 1997; Szabo & Bak, 1999; Szabo, Mesko, Caputo & Gill, 1998; Treasure & Newbery, 1998). In the first review to focus on the measurement of affect in the context of exercise, Gauvin and Spence (1998) noted that, although a number of questions regarding the EFI remain to be investigated, the preliminary indications regarding the validity and reliability of the scale are promising. An independent examination of the factor structure of the scale also led to generally positive conclusions (Vlachopoulos, Biddle & Fox, 1996).

The preliminary evaluations of the EFI, however, have focused exclusively on the computation of psychometric indices and have failed to provide a critical assessment of the conceptual underpinnings of the scale and the methods that were employed in its development. This is an important omission, especially given the fact that the EFI introduces a novel construct with unique structure. Evaluating whether the structure of a data set is consistent with a proposed model without considering whether the model was appropriately conceptualized in the first place seems problematic. Statistical fit has no bearing on whether the model itself is theoretically meaningful. The theoretical merit must be evaluated independently and prior to any psychometric assessments. It is also noteworthy that the researchers who have used the EFI in applied studies have not cited any conceptual reasons as a basis for selecting this particular instrument as opposed to some other relevant measure.

In summary, the EFI has yet to undergo a critical and thorough evaluation. Given its growing popularity, it is essential that such an evaluation be performed. To this effect, the present analysis concentrates on the conceptual foundations of the EFI and the methodological steps that were followed in its development. In both areas, the theoretical and methodological literatures on affective phenomena have served as guides. Thus, the discussion of the conceptual issues is embedded within the broader framework of affect theorizing and the methodological steps that were followed in the development of the EFI are contrasted with established guidelines.

Conceptual foundation

Delineation of content domain

As its name declares, the EFI (Gauvin & Rejeski, 1993) was developed as a measure of exercise-induced "feeling." It is important to note that the term "feeling" has seldom been used in the affect literature, mainly to avoid the potential for confusion that its multiple connotations entail. Notably, Averill (1994) has characterized "feeling" as "one of the vaguest terms in the

English language" (p. 379). Lazarus (1991) also noted the ambiguity inherent in the term and suggested that it be used only in reference to "the awareness of bodily sensations" (p. 57).

A precise definition is necessary for the development of any measure, but particularly in the case of a measure of "feeling," this first step seems absolutely essential. However, a direct definition of the term "feeling" is absent from the original publication by Gauvin and Rejeski (1993). As an indirect clue, the intended content domain of the EFI was described as encompassing "those states that are directly tied to the stimulus properties of physical activity" (p. 405) and the "feelings that evolve from the intrinsic qualities of the exercise experience" (p. 406). In an effort to define "feeling states" directly, Gauvin and Spence (1998) later stated that the term refers to "those human experiences that include bodily reactions, cognitive appraisals, actual or potential instrumental responses, or some combination thereof" (p. 326). However, consistent with Averill's (1994) characterization of the term as "one of the vaguest," the inclusion of bodily reactions, cognitive appraisals, and instrumental responses in Gauvin and Spence's (1998) definition creates an overly inclusive scope, blurring the exact nature and the limits of the content domain of the EFI. The absence of a tight definition is crucial, as it can only exacerbate the confusion and complicate the validation process.

Theorized structure of content domain: categories or dimensions?

The EFI was developed as a measure of the general domain of "exercise-induced feeling" and yet it only targets four distinct states. Specifically, Gauvin and Rejeski (1993) contended that "the stimulus properties of physical activity are capable of producing several distinct feeling states" (p. 406, emphasis added), namely liveliness, tranquility, enjoyment, and fatigue (note that the subscales of the EFI that correspond to these constructs were labeled slightly differently: Revitalization, Tranquility, Positive Engagement, and Physical Exhaustion, respectively). The focus on these four states was characterized as an "a priori conceptual framework" (p. 411).

Thus, although not described in these terms, "exercise-induced feeling" was viewed as a categorical, as opposed to a dimensional domain. As we have explained previously (Ekkekakis & Petruzzello, 2000), it is generally acknowledged that categorical models are useful primarily when the focus of investigation is on deciphering the distinct antecedents and experiential characteristics of specific emotions (i.e., those affective states that are elicited following various patterns of cognitive appraisal). On the other hand, dimensional models, because of their wider scope and the potential for a parsimonious representation of the global affective space, are considered particularly useful when the purpose of an investigation is to provide a general description of the nature and the dynamics of affective responses elicited in a given context.

Until now, most studies of the exercise-affect relationship have followed the categorical approach, focusing on distinct affective states, such as state anxiety, depression, or various mood states. This approach has allowed researchers to determine whether exercise is associated with changes in these particular states, but can offer no indication as to whether these changes are the primary or most experientially salient affective changes that take place. In other words, although it has been possible to determine that, under certain exercise conditions, people are likely to feel, for example, less anxious and more vigorous, this does not mean that this is all that people feel at the time or that these are the most salient of the affective changes that they experience. These questions cannot realistically be addressed by examining distinct affective states. Instead, these

questions necessitate a model with a wide, theoretically unrestricted, scope. To the best of our knowledge, except for a few isolated attempts (e.g., Kinsman & Weiser, 1976; Morris & Salmon, 1994), a systematic investigation of how different people feel during and following various exercise conditions has yet to be conducted.

Therefore, given (a) how little is presently known about the kinds of affective responses that are likely to emerge under various exercise conditions and (b) that the main strength attributed to dimensional models of affect is their ability to provide a broad, balanced, and parsimonious representation of the global affective space, it seems reasonable to suggest that investigations aimed to provide a general description of how people feel during and following exercise would benefit more from a dimensional, rather than a categorical, approach. From this perspective, Gauvin and Rejeski's (1993) decision to develop the EFI as a categorical measure is puzzling. To further complicate matters, in a very informative chapter co-authored by one of the developers of the EFI and published in the same year as the EFI (Gauvin & Brawley, 1993), an eloquent case was made for using a dimensional, as opposed to a categorical, approach for studying affect in the context of exercise:

> [The dimensional] approach seems better suited to the understanding of exercise and affect because the models stemming from it are intended to be broad, encompassing conceptualizations of affective experience. Because the affective experience that accompanies exercise has not been thoroughly described, a model of affect that has a wider breadth is more likely to capture the essence of exercise-induced affect than a model that, at the outset, limits the focus of investigation to specific emotions (p. 152).

This statement, which is consistent with our interpretation of the relative advantages of dimensional models over categorical ones, raises some obvious questions: If dimensional models are deemed preferable for use in the context of exercise (and for aptly articulated reasons), then why was the EFI developed as a categorical instrument? Given that "the affective experience that accompanies exercise has not been thoroughly described," why run the risk of "limiting the focus of investigation" to specific affective states? These are important, albeit unanswered, questions.

Components of exercise-induced feeling

As Cronbach and Meehl (1955) stressed almost a half century ago, a necessary precondition for construct validation is to build an adequate "nomological net" that specifies the unique features of the constructs of interest and their conceptual ties to related constructs. The construction of an adequate nomological net in the case of the EFI would require (a) the precise definition of the constructs of liveliness, tranquility, enjoyment, and fatigue, (b) a description of their distinct antecedents and unique links to exercise, (c) a postulated model of their inter-relationships, (d) a rationale for delimiting the content domain of the EFI to these particular affective constructs to the exclusion of others, and (e) a delineation of the relationships of these constructs to other affective constructs. However, most of this information is absent from the material provided by Gauvin and Rejeski (1993).

In fact, the rationale for focusing on liveliness, tranquility, enjoyment, and fatigue was outlined in only two paragraphs. This limited coverage is not surprising, considering the complete lack of

information on the kinds of affective responses that may emerge among different people under different exercise conditions. Consequently, the argumentation consisted primarily of conjectures. First, it was proposed that physical activity is linked to liveliness. It was argued that, contrary to increases in sympathetic activity that accompany stress, which are typically associated with tension, increases in sympathetic activity associated with physical activity "can produce feelings of being refreshed or alive" (Gauvin & Rejeski, 1993, p. 405). Second, it was proposed that physical activity is linked to tranquility due to findings of reduced electromyographic activity and increased alpha power in the electroencephalograph associated with exercise. Third, it was proposed that physical activity is linked to enjoyment. This is because of (a) epidemiologic data indicating that physical activity is associated with more happiness compared to recreational activities and household chores and (b) experimental data indicating that exercise is associated with decreases in anxiety and tension, an effect that is attenuated when exercise is combined with exposure to mental stressors. It is noteworthy that even Gauvin and Rejeski themselves questioned the generalizability of the enjoyment-producing effects of exercise, stating that "the arousal created by physical activity can lead to a variety of feelings, good or bad, depending upon the cognitive label assigned to events" (p. 405, emphasis added). Fourth, it was proposed that physical activity is linked to fatigue due to "thermal discomfort, the depletion of metabolic resources, respiratory distress, and sensations originating from the muscles and joints" (Gauvin & Rejeski, 1993, p. 405).

Although these affective states may, in fact, be influenced by exercise under certain conditions, the arguments presented by Gauvin and Rejeski (1993) in support of the proposed links were, by necessity, mainly speculative and, as such, debatable. For instance, most theorists would question the postulated connection between sympathetic activation and perceived liveliness and the connection between electromyographic or electroencephalographic activity and perceived tranquility (see Cacioppo & Tassinary, 1990, for an insightful explication of the difficulties inherent in inferring psychological constructs from physiological events). Moreover, no arguments were presented to support the decision to focus on liveliness, tranquility, enjoyment, and fatigue to the exclusion of all other "feeling states" that may be influenced by exercise.

On the notion of a global phenomenology of exercise: conceptual problems

Gauvin and Rejeski (1993) asserted that the "universe of content" of the EFI "is consistent with the phenomenology of men and women who exercise with some regularity" (p. 419) and, more generally, that the components of the EFI reflect the "phenomenology of people involved in exercise in the real world" (p. 408). More recently, Rejeski, Reboussin, Dunn, King and Sallis (1999) went further, stating that the content of the EFI reflects the "primary forms of affect that are directly influenced by physical activity" (p. 98).

These statements raise a contentious issue, which is made even more vexing by the fact that the statements were based on little or no empirical evidence. The notion that one can identify the "phenomenology of people involved in exercise in the real world" and adequately capture it in a small and invariant set of psychometric scales is tantamount to a claim of a universal phenomenology of exercise — as if there were only one exercise and only one exerciser. Contrary to this idea, it seems reasonable to suggest that liveliness, tranquility, enjoyment, and fatigue would not represent an accurate and comprehensive account of what sedentary, unfit, obese, elderly, injured, diseased, or otherwise physically limited individuals experience during exercise. As one example,

when McAuley, Peña, Katula and Talbot (1997) examined the responses of older adults (mean age 66 years) to a graded exercise test, they noted an "overwhelming negative response" which extended beyond an increase in fatigue. Likewise, using the Activation Deactivation Adjective Check List (AD ACL; Thayer, 1989), Ekkekakis, Hall and Petruzzello (1999) found significant increases in tension in response to a 30-min bout of treadmill running at 75% of maximal aerobic capacity. These findings demonstrate that, depending on the characteristics of the exercisers and the attributes of the exercise stimulus, exercise may acquire very diverse and variable "stimulus properties" that are likely to fall outside the scope of the EFI.

This is hardly a controversial claim. In other papers published around 1993, Gauvin and Rejeski themselves expressed strong opposition to nomothetic conceptions of the exercise-affect relationship (Gauvin & Brawley, 1993; Rejeski, 1994; Rejeski & Thompson, 1993). Furthermore, Gauvin and Spence (1998) admitted that the four scales of the EFI do not fully capture the "phenomenology of people involved in exercise." They stated that "the stimulus properties of exercise can elicit feeling states including **but not limited to** positive engagement, revitalization, physical exhaustion, and tranquility" (p. 327, emphasis added). This statement is perplexing. If there is more to "exercise-induced feeling states" than "positive engagement, revitalization, physical exhaustion, and tranquility" then what was omitted from the EFI and why? For instance, what is one to make of the absence of any unequivocally negatively valenced scales (fatigue can be, but is not necessarily negatively valenced)? Does the EFI only arbitrarily capture the positive phenomenology of exercise or perhaps even only some part of the positive phenomenology of exercise? Does the exclusion of negatively valenced affective states imply that "the stimulus properties of physical activity" are not tied to any negatively valenced states? Or does it imply that, although such states do occur, they were excluded from the EFI because they were considered unimportant, secondary, or only indirectly tied to "the stimulus properties of physical activity?"

These issues have important practical implications. Consider, for instance, a case in which a researcher takes for granted the claim that the EFI can provide an assessment of the "phenomenology of people involved in exercise in the real world" and the "primary forms of affect that are directly influenced by physical activity." Resting on this assumption, one would consider it unnecessary to embark on the extremely laborious enterprise of independently investigating other possibly relevant aspects of this "phenomenology" on a case-by-case basis. Consequently, the EFI would be used as the sole measure of affect and all conclusions about "exercise-induced feeling states" would be based exclusively on observed changes in the four subscales of the EFI. If exercise failed to induce significant changes in these variables, but, instead, exercisers felt more tense or bored (or experienced emotional responses like anger or embarrassment), the researcher would conclude, erroneously, that exercise under the given experimental conditions produced no significant changes in "feeling states" in general.

This is not a fictional scenario. The EFI is being used routinely in applied research as the sole measure of affect (e.g., Annesi & Mazas, 1997; Turner, Rejeski & Brawley, 1997; Vlachopoulos, Biddle and Fox, 1996), even in cases of people explicitly selected to be sedentary (i.e., that do not "exercise with some regularity"; for example, see Gauvin, Rejeski, Norris & Lutes, 1997; Treasure & Newbery, 1998). In such cases, the scope of the EFI makes it impossible to infer what, if any, affective changes actually occurred. To illustrate the point, in a study of adult sedentary community dwellers, Gauvin et al. (1997) could only comment on the absence of "widespread mood enhancing effects of acute exercise" and the absence of a "strong dose-response

relationship" (p. 509). Of course, it is entirely possible that changes in "feeling" or affective states other than those assessed by the EFI actually did occur. Likewise, it is possible that a dose-response relationship did emerge, but was reflected in "feeling" or affective states other than those assessed by the EFI. These possibilities challenge the notion of a universal phenomenology of exercise.

On the notion of a measure of exercise-specific affect: logical problems

Gauvin and Russell (1993) have cautioned that devising measures that are specific to sport or exercise situations should not be done uncritically or in the absence of a theoretical framework. In some cases, domain-specificity may be unwarranted or downright inappropriate. The study of the exercise-affect relationship may be one such case. A measure of affect tailored to tap only those affective states believed to be relevant to exercise leads to some considerable logical problems. An important problem stems from the fact that most studies investigating the exercise-affect relationship involve assessments of affect under non-exercise conditions, such as before exercise or during waiting list control conditions. If the scope of the measure has been restricted a priori to tap only those states likely to be influenced by exercise (i.e., the so-called "exercise-induced feeling states") and exclude all other variants of affective experience, then how meaningful would any comparisons be between exercise and all the non-exercise conditions where this measure is likely to be employed? Could a person ever feel *refreshed* or *revived* (items from the EFI) under conditions of quiet rest, given that by their very grammatical form (i.e., past participle) and meaning these items assume an ongoing or preceding relevant stimulus, such as exercise? Would the comparison between responses to exercise and non-exercise conditions in these items constitute a compelling demonstration of affective benefits associated with exercise? Or would the findings instead be confounded by the fact that *revived* was specifically chosen because of its possible relevance to exercise and not to sitting quietly?

Besides the logical problems, given that the items of the EFI were specifically selected to be responsive to exercise, but not to non-exercise conditions, when the instrument is used in non-exercise conditions, a host of statistical problems are also likely to arise. Specifically, mean ratings in non-exercise conditions are likely to remain very close to the bottom of the possible range, leading to problems typically described in psychometrics as "floor effects." Means near the bottom of the range are commonly associated with violations of the assumption of normality, which is essential for most statistical tests. This may also have a suppressive effect on item variances. In turn, low variances entail low covariances, which may manifest themselves as reduced internal consistencies and collapsed factor structures, thus also leading to violations of the assumption of factorial invariance across exercise versus non-exercise conditions. Again, despite these possible problems, the EFI is being used routinely in conjunction with non-exercise control conditions (e.g., Bozoian, Rejeski, & McAuley, 1994; Gauvin, Rejeski & Norris, 1996; Gauvin et al., 1997; Rejeski, Gauvin, Hobson & Norris, 1995; Treasure & Newbery, 1998; Turner et al., 1997).

Conclusion

A critical examination of the conceptual underpinnings of the EFI raises several concerns. These are mainly related to deficiencies in conceptual groundwork that can potentially undermine

the construct validity of the EFI by weakening the nomological net (Cronbach & Meehl, 1955) on which the development of the scale was based. This complicates the evaluation of the methodological steps involved in the development and validation of the EFI.

Methodological implementation of conceptual precepts

This section focuses on the most fundamental stages in the development of the EFI, namely the selection of items and the structural analyses. The methods that were followed are contrasted with traditional guidelines for the development of psychometric instruments (Anastasi, 1990; Carmines & Zeller, 1979; Clark & Watson, 1995; Crocker & Algina, 1986; Cronbach, 1990; DeVellis, 1991; Nunnally & Bernstein, 1994). Each methodological decision is examined from a conceptual and a technical-psychometric standpoint and its possible impact on the content and structure of the EFI is highlighted.

Overview

In general, several of the methodological steps that were taken in the development of the EFI deviated considerably from established procedures. Before discussing these deviations and their possible implications, it is useful to review some basic guidelines for the development of psychometric instruments. According to the "Standards for Educational and Psychological Testing" of the American Psychological Association (APA, 1985), "the first task for test developers is to specify adequately the universe of content that a test is intended to represent" (p. 10). DeVellis (1991) also emphasized that "the boundaries of the phenomenon must be recognized so that the content of the scale does not inadvertently drift into unintended domains" (p. 51). Once the domain of content has been precisely defined and demarcated, the process of item selection or generation may begin, using expert judges or development samples to establish content validity (i.e., to ensure that the intended content domain has been targeted successfully by the item pool; see Carmines & Zeller, 1979; Crocker & Algina, 1986; Haynes, Richard & Kubany, 1995).

At this point, scale developers may follow one of two routes, depending on the presence or the absence of an adequate theoretical basis for the structure of the content domain. The deductive route is followed when there is a sufficient basis for enunciating specific a priori postulates about the number, the nature, and the inter-relationships of the components of the domain. This foundation may be built either by conceptual reasoning or by previous empirical study. In a nutshell, the defining characteristic of a deductive strategy is that the "choice and definition of constructs precede and govern the formulation of items" (Burisch, 1984, p. 215). Alternatively, in the absence of a full-fledged theoretical basis, an inductive approach may be followed. In this case, "one starts with a collection of individual items, lets the data 'speak for themselves', and ends up with scales at a higher level of abstraction" (Burisch, 1984, p. 215). In other words, the defining characteristic of an inductive strategy is that "the number and nature of the resulting scales follow from the data analysis" (Burisch, 1984, p. 215).

It must be emphasized that, although an inductive strategy is certainly an acceptable means of approaching a novel domain, certain rules must be followed. First of all, an inductive approach is appropriate only when the structure of the content domain is not fully known a priori, not in

the complete absence of a theoretical framework. The presence of specific and concrete theoretical postulates about the nature and the limits of the content domain is sine qua non. According to the APA (1985), "the construct of interest for a particular test should be embedded in a conceptual framework, no matter how imperfect that framework may be. The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables" (pp. 9-10). A similar point was made by DeVellis (1991): "Even if there is no available theory to guide the investigators, they must lay out their own conceptual formulations prior to trying to operationalize them" (p. 52).

What these guidelines guarantee is that the nature and the boundaries of the construct to be measured (i.e., the content domain) will not be altered in the process of exploring the structure of the construct. Reshaping the content domain itself in the process of scale development is not acceptable because it undermines the logic of the entire validation process. Simply, if the content domain itself continuously changes, which of its transformations should be viewed as the standard against which the validity of the measure is to be evaluated?

When employing an inductive strategy, the use of empirical methods is essential (Loevinger, 1957). Relying on intuition or subjective judgement for devising scales (i.e., what is often referred to as the "rational" approach) exposes the scale development process to a host of potential biases. Nunnally and Bernstein (1994) have characterized rational approaches, as opposed to approaches based on empirical methods and quantitative psychometric criteria, as "fundamentally inadequate" (p. 319). To repeat Burisch's (1984) phrase, scale developers must "let the data speak for themselves". To that effect, they have at their disposal an extensive array of psychometric techniques. Commenting on the advantages associated with the use of factor analysis for this purpose, Gorsuch (1997) remarked that this technique, "instead of basing the factors on investigator judgement, it bases each factor on a set of highly correlated items. Hence, misjudgements about what items measure are less likely to distort the operationalization of the construct" (p. 535).

Finally, since some items are likely to be better indices of a certain component than other items (i.e., exhibit a higher degree of "content saturation" or higher reliability), an additional step may be necessary, to select the psychometrically stronger and discard the psychometrically weaker items. Again, to minimize the possibility of biases influencing this process, this screening should be made "according to the properties of the items revealed **by empirical study**" (Loevinger, 1957, p. 658, emphasis added).

Development, item selection, and content validation

Gauvin and Rejeski's (1993) claim that the focus on liveliness, tranquility, enjoyment, and fatigue was prompted by an "a priori conceptual framework" (p. 411) pointed to a deductive approach to scale development. Assuming a deductive approach, the reasonable first step would have been the sampling of items that were judged to be valid indices of the four specific constructs of interest. This judgement should have been based on standard procedures (i.e., use of expert judges and/or development samples). However, instead of sampling items representative of liveliness, tranquility, enjoyment, and fatigue, items for the initial pool were sampled indiscriminately from the entire affective lexicon (resulting in over 500 items). Likewise, instead of using expert judges and development samples for the purpose of content-validating items as indices of liveli-

ness, tranquility, enjoyment, and fatigue, the only screening criterion that was employed was the relevance or lack thereof of items to exercise.

Given what was actually done, the purpose of this initial phase evidently was not to select the best indices of four specific affective constructs. Instead, it appears that the purpose was to identify affective states that may arise in the context of exercise, in an effort to explore the controversial idea of a universal phenomenology of exercise (see previous section for a critique). Thus, the initial item pool was reduced by asking three investigators in exercise psychology to evaluate each item on whether the affective state it described "occurs" during or following exercise or not. The 145 items that were retained were then shown to a sample of physically active college students, who were asked to confirm that the affective states they described were, in fact, relevant to "the effects that preceded or followed either low-intensity or high-intensity exercise" (p. 407). All 145 items were retained.

Given that the purpose of this process was to explore the "phenomenology of exercise," some limitations are worth noting. Specifically, it is clear that the phenomenological profile that could emerge from these procedures would have limited generalizability. In essence, the judgement of relevance to exercise was based only on the opinions of three experts, since the students, who presumably had a broader range of exercise experiences, did not have the opportunity to express their experiences in an open-ended fashion. Yet, in spite of this problem, Gauvin and Rejeski (1993) interpreted the agreement between experts and students as evidence that "the students' phenomenology essentially paralleled the judgements of the experts" (p. 407). It is also important to note that, even if the students were allowed to report their experiences and substantively contribute to the item selection process, their young age and physically active status would constitute a significant biasing factor. The resultant phenomenological profile would still only reflect the experiences of a small portion of the population. It is reasonable to assume (and open to empirical verification) that the study of a wider range of populations (e.g., sedentary, elderly, injured, diseased, or otherwise physically limited individuals) would have resulted in a different profile. Again, in spite of these limitations, Gauvin and Rejeski argued that the content of the EFI reflects the "phenomenology of people involved in exercise in the real world" (p. 408) and Rejeski et al. (1999) maintained that it accounts for the "primary forms of affect that are directly influenced by physical activity" (p. 98). As noted previously, the issue of generalizability is crucial, particularly given the fact that the EFI is commonly used as the sole measure of affect, even in studies involving individuals explicitly selected to be sedentary (e.g., Gauvin et al., 1997; Treasure & Newbery, 1998).

Further reduction of item pool and determination of structure of content domain

Following the initial screening, instead of having a manageable pool of content-validated items (i.e., items judged as valid indices of the components of the content domain), in the case of the EFI, what remained was a large pool of 145 items that were deemed relevant to the context of exercise but reflected a variety of affective constructs. This posed a challenge, as a large and diverse (i.e., not content-validated) item pool had to be organized and reduced to a manageable size. Doing so by means of an exploratory factor analysis would have been impractical, as it would have required an inordinate number of participants.

The solution that was devised to address this problem was atypical of psychometric procedures

for two reasons. It confounded substantive and structural considerations and it relied on semantic analyses and subjective criteria. Specifically, Gauvin and Rejeski (1993) initially performed a qualitative content analysis aimed at grouping the 145 items into clusters. This procedure resulted in the formation of 15 such clusters. Then, Gauvin and Rejeski eliminated 11 of the 15 clusters in two steps (7 in the first, 4 in the second). The four remaining clusters (of 3 items each) were considered representative of liveliness, tranquility, enjoyment, and fatigue, respectively. Thus, the final form of the EFI comprised 12 items organized in 4 scales.

Given the impact of these procedures on the final content of the EFI, further scrutiny is warranted. First, it is essential to understand the classification rules that were followed in the formation of the 15 item clusters. However, these were not disclosed. The only information that was provided by Gauvin and Rejeski (1993) was that the items were grouped in "conceptually homogeneous categories based on previous dimensional analyses of affect-laden words" (p. 407) and cited four sources as conceptual guides in their analysis (i.e., Frijda, 1987; Izard, 1977; Ortony, Clore & Foss, 1987; Russell, 1980). The use of the terms *categories* and *dimensional* in this statement is puzzling, since categorical and dimensional models are antithetical in most of their underlying principles. The confusion is exacerbated by examining the content of the cited sources, since they, too, constitute a mix of categorical and dimensional models. Specifically, Frijda (1987) and Izard (1977) are proponents of the idea of basic emotions (i.e., a variant of the categorical approach), but differ in terms of which emotions they consider basic. Ortony et al. (1987) also support the notion of categories, but are known critics of the idea of basic emotions (see Ortony & Turner, 1990). Finally, Russell (1980) is the only theorist of those mentioned who has dealt with affect in general (i.e., not exclusively with emotions), but in his 1980 work that Gauvin and Rejeski cited, he proposed a dimensional model (i.e., the affect circumplex), which goes against the idea that affective states can be grouped in distinct categories. Therefore, given that the sources cited as guides are conceptually disparate and that the models they propose do not have any apparent overlap with the 15-cluster categorization adopted by Gauvin and Rejeski (1993), it is ultimately impossible to infer what classification rules were applied. This makes the procedure impossible to replicate.

The subsequent process of eliminating 11 of the 15 clusters is also critical. However, again, the criteria that were used were not adequately documented. The following arguments were presented as a basis for the initial elimination of seven clusters: (a) those clusters were perceived as representing consequences (i.e., secondary manifestations) of the constructs represented by the remaining eight clusters which were evidently perceived as primary (for example, alertness and depression were considered dependent upon activation and revitalization; anxiety was considered dependent upon tranquility), (b) the constructs represented by some of the eliminated clusters could be assessed by existing psychometric instruments, such as the POMS, (c) clusters referring to physical symptoms were eliminated because "physical symptoms tend to be highly variable across subjects" (Gauvin & Rejeski, 1993, p. 407), and (d) clusters of items referring to potency or disappointment were dropped because they were considered "more relevant to the study of chronic training effects" (p. 407). It is important to point out that these claims were made in the absence of any empirical or theoretical evidence and the rationale behind some is enigmatic. For example, it is unclear why the fact that a certain construct is also assessed by another instrument (e.g., the POMS) would be grounds for disqualifying this construct from inclusion. This decision should be strictly a function of whether the construct is part of the content domain as defined and demarcated.

Four additional clusters were eliminated based on (a) "preliminary factor analyses" (Gauvin & Rejeski, 1993), (b) an "awareness that the first eight clusters could be adequately represented by four" (p. 408), (c) linguistic redundancy, and (d) concerns regarding lack of face validity in some items. Again, no further details were provided because Gauvin and Rejeski considered that it would have been "unnecessarily cumbersome to present the analytical details of these procedures" (p. 423). However, given the impact of these decisions on the structure of the EFI (i.e., the elimination of half of the remaining clusters), the non-disclosure of this information leaves a critical void and makes it impossible to replicate these procedures.

In summary, the substantial deviations from established scale development guidelines hinders the process of evaluation. It is clear that the development of the EFI was not based on a deductive process. At the same time, the appropriate rules for an inductive approach were not followed. Had the appropriate procedures been followed, Gauvin and Rejeski should have "let the data speak for themselves" (Burisch, 1984, p. 215) by examining the structure of the item pool and the psychometric merit of the items using standard quantitative methods (i.e., item analyses, factor-analytic techniques, etc.). The content validity of the 12 items that remained was never formally examined. Given that the final content of the EFI was molded during the process of item selection mainly on the basis of subjective and inadequately documented criteria, no definitive statements can be made about the nature, the boundaries, the representativeness, or the structure that this content represents.

Factor structure

Exploratory factor analytic techniques are typically used in the process of scale development either to investigate the structure of the content represented by the item pool (when the structure is unknown and there are insufficient theoretical grounds to postulate what it would be) or to identify the items that are the strongest indices of the constructs of interest (Briggs & Cheek, 1986; Clark & Watson, 1995; Comrey, 1988; Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1997; Nunnally & Bernstein, 1994). In the case of the EFI, neither purpose appears relevant. First, the 12 remaining items were presented as having a known or postulated dimensionality (i.e., they were already categorized in clusters labeled Revitalization, Tranquility, Positive Engagement, and Physical Exhaustion). Second, the identification of strong items to be retained and weak items to be discarded was impossible, since there were only three items available per hypothesized scale.

Nevertheless, the 12 items that remained from the previous procedures were subjected to a principal components analysis. The various aspects of this analysis are reviewed here and compared to established guidelines (Comrey & Lee, 1992; Fabrigar, Wegener, MacCallum & Strahan, 1999; Ford, MacCallum & Tait, 1986; Gorsuch, 1983; Kim & Mueller, 1978).

In conducting the analysis, Gauvin and Rejeski (1993) used the options that most statistical software packages use by default. This practice, however, is known to often result in choices that are inappropriate (Fabrigar et al., 1999; Gorsuch, 1997; MacCallum, 1983). First, let us examine the appropriateness of the model of factor analysis that was used, namely principal components analysis. Although Gauvin and Rejeski used the terms *principal component* and *factor* interchangeably, principal components analysis (PCA) and common factor analysis (FA) are distinct procedures, used for different purposes (Millsar & Meredith, 1992). PCA is used as a data

reduction method. Its purpose is to use a small number of linear transformations of items to account for as much variance in those items as possible (including error variance). Components should not be considered reflective of underlying latent constructs. On the other hand, FA is aimed at explaining only common item variance (i.e., variance shared by the items). Simply put, "PCA analyzes variance; FA analyzes covariance" (Tabachnick & Fidell, 1996, p. 663). In contrast to PCA, the covariation found among items in FA can be attributed to latent variables (Floyd & Widaman, 1995; Kim & Mueller, 1978; Widaman, 1993). Therefore, when the purpose of the analysis is not to simply reduce an item pool, but to seek commonalities that can be attributed to underlying latent constructs, the appropriate analytic model is FA, not PCA.

Until a few years ago, most researchers would probably consider the practical implications of the distinction between PCA and FA trivial and, in the presence of high communalities (i.e., above .70) and a large number of items under analysis, this may actually be the case (Velicer & Jackson, 1990). Although Gauvin and Rejeski (1993) did not report communalities, these, if calculated, can be shown to be relatively high (all higher than 0.62). However, Monte Carlo studies have demonstrated that, with a small number of items in the analysis (only 12 in the case of the EFI) (Snook & Gorsuch, 1989) and, more specifically, a small number of items per component (only three in the case of the EFI) (Widaman, 1993), PCA can lead to substantial inflation of item loadings. Widaman (1993) estimated that, with three items per component, PCA can inflate the magnitude of loadings by 0.157 for population loadings of 0.60 or 0.263 for population loadings of 0.40. Discussing these findings, Snook and Gorsuch (1989) noted the following: "The inflated component loadings were generally large enough to affect how data are interpreted. Because the loadings were inflated, the results would appear 'clearer' to the user, even though they would be misleading. This false clarity may be one reason why component analysis is more popular than common factor analysis in applied research" (p. 151).

Second, Gauvin and Rejeski (1993) only cited using the "eigenvalue greater than 1" criterion, also known as the Guttman-Kaiser (GK) rule (Guttman, 1954; Kaiser, 1960, 1970), as a basis for deciding how many components to retain. The reasoning behind the GK rule is that a component with an eigenvalue less than 1 would account for less variance than a single variable and would also be unreliable (Cliff, 1988; Comrey, 1978). However, in Monte Carlo studies, the GK rule has consistently been shown to be the least accurate of all indices of dimensionality, leading to overextraction in most cases (Cattell & Vogelmann, 1977; Hakstian, Rogers & Cattell, 1982; Revelle & Rocklin, 1979; Zwick & Velicer, 1982, 1986). Specifically, the number of components indicated by the GK rule tends to be 1/3 to 1/5 of the total number of items being examined, regardless of the actual number of components (Velicer & Jackson, 1990; Zwick & Velicer, 1982, 1986). Furthermore, Cliff (1988) demonstrated that, contrary to popular belief, the GK rule does not show how many components will be reliable. Therefore, it is generally agreed that the GK rule should not be relied upon for deciding how many components to retain. This has potentially significant implications for the EFI. Based on the GK rule, Gauvin and Rejeski (1993) initially extracted three principal components. This is consistent with the aforementioned Monte Carlo findings showing that, on average, the GK rule leads to the extraction of a number of components that is 1/3 to 1/5 of the total number of items in the analysis, which in most cases represents an overestimate. From this perspective, it is useful to examine the scree plot (Cattell, 1966) of the first four eigenvalues from Gauvin and Rejeski's PCA. The scree test has generally been shown to be a more accurate index of dimensionality compared to the GK rule (Cattell & Vogelmann,

1977; Zwick & Velicer, 1982, 1986). As can be seen in Fig. 1, a clear elbow appears at the second eigenvalue, suggesting only two components. This piece of information is potentially critical, given that, as will be discussed later, Gauvin and Rejeski eventually decided to retain not two (viz. the scree test), not three (viz. the GK rule), but four components.

Third, Gauvin and Rejeski (1993) followed the PCA with both orthogonal and oblique rotations. However, the method of rotation should be a conceptually informed decision. According to Pedhazur and Schmelkin (1991), "from the perspective of construct validation, the decision whether to rotate factors orthogonally or obliquely reflects one's conception regarding the structure of the construct under consideration. It boils down to the question: Are aspects of a postulated multidimensional construct intercorrelated?" (p. 615). If Gauvin and Rejeski had postulated that liveliness, tranquility, enjoyment, and fatigue were related, an orthogonal rotation would have been inappropriate. On the other hand, if they had postulated that the constructs were unrelated, they should have cited this premise as a basis for conducting an orthogonal rotation. In both cases, they should have examined the configuration of the components graphically to ensure that the components were properly formed and that their choice of rotation was appropriate. No such considerations were cited by Gauvin and Rejeski. However, given the conceptual links between liveliness and enjoyment and the semantic similarities between the items that were selected to represent these constructs, it seems reasonable to suggest that the corresponding components would be correlated. In that case, an oblique rather than an orthogonal rotation would have been more appropriate.

Although these analytic choices appear to be inconsistent with psychometric principles, their impact on the structure of the EFI cannot be directly assessed from the data reported by Gauvin and Rejeski (1993). However, a number of important observations can be made from the data that were reported. We have already commented on the decision to retain a 4-component structure. As we pointed out, the GK rule showed that three components should be extracted, whereas the generally more accurate scree test indicated only two. Given the evidence that the GK rule tends

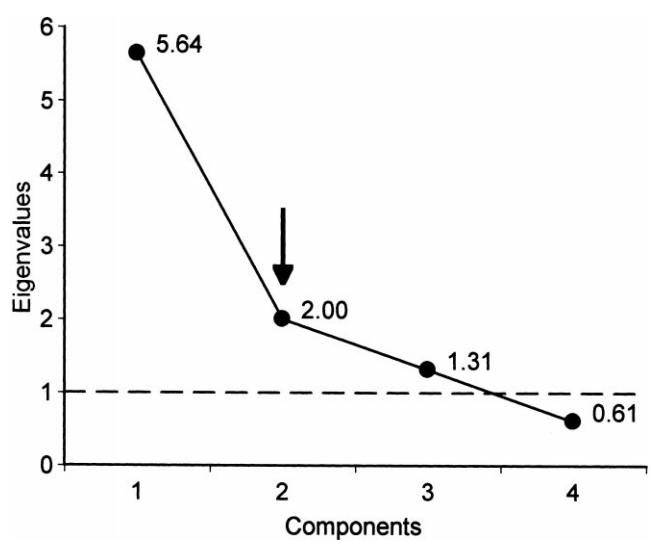

Fig. 1. Scree plot of the eigenvalues of EFI components.

to overestimate the number of components, the decision should have been made between two or three components, but not more.

This conclusion can be supported by additional evidence. First, the size of the fourth eigenvalue was only 0.61, over 9 times smaller than the first. It is true that components with eigenvalues very close to 1 should not be discarded without additional evidence (Schonemann, 1990), but the size of this eigenvalue indicates that the fourth component could only account for a little over half of the variance contributed by a single variable, so retaining it does not make sense from a factor analytic perspective.

Second, the 4-component solution lacked simple structure (Veldman, 1974), a term referring to the idea that each variable should load highly on only one component (Kline, 1994; Merenda, 1997; Nunnally & Bernstein, 1994). Complex structures cannot be meaningfully interpreted. In the case of the EFI, the 3-component solution led to a fairly well-defined simple structure after varimax rotation, whereas the 4-component solution led to a visible deterioration of simple structure in the third and fourth components. For example, the item *enthusiastic* had loadings of 0.59 on the third component and 0.64 on the fourth, and the item *up-beat* had loadings of 0.57 on the third component and 0.53 on the fourth. Given these cross-loadings, a distinction between these items is unjustified on statistical grounds.

Third, evidence for the inappropriateness of the 4-component structure can be obtained by examining the pairwise component plots. As can be seen in Fig. 2a and b from the 3-component solution, the items *enthusiastic*, *up-beat*, *happy*, *energetic*, *refreshed*, and *revived* are grouped in a relatively tight cluster. On the contrary, in Fig. 2c from the 4-component solution, the complete deterioration of simple structure and the absence of a clustering pattern among the six items is obvious. It is important to emphasize that the plots do not show that the structure was oblique (i.e., that the components were non-orthogonal or correlated), but rather that no true components were formed. Psychometricians strongly caution researchers to recognize and avoid such occurrences (Kim & Mueller, 1978; Nunnally & Bernstein, 1994). According to Kline (1994), the failure to obtain simple structure is the main cause of "confusion and misleading results in exploratory analyses" (p. 182) and advocating the extraction of components without identifying them in factor space and without reference to external criteria is "one of the easiest roads to delusion" (p. 181) in test construction. What happened in the 4-component solution of the EFI was a case of "splitting" or "fission", a common consequence of overextraction (Wood, Tataryn & Gorsuch, 1996).

Despite these problems, Gauvin and Rejeski (1993) labeled the 4-component solution a "conceptually driven structure" (p. 410) and decided to retain it for further analysis. Specifically, they proposed that the items *enthusiastic*, *happy*, and *up-beat* formed a component labeled Positive Engagement and the items *energetic*, *refreshed*, and *revived* formed a separate component labeled Revitalization.

Gauvin and Rejeski (1993) presented three arguments to justify their decision to circumvent the results of the PCA. First, they stated the following: "Revitalization and Positive Engagement may not represent distinct feeling states. Based on our conceptual structure, however, we think that this explanation is unlikely" (Gauvin & Rejeski, 1993). Yet, as previously noted, the exact nature and the theoretical underpinnings of this "conceptual structure" were unclear. Of the sources that Gauvin and Rejeski cited as the basis of their earlier item classification, Russell's (1980) model was the only one that dealt with affect in general (i.e., was not limited to emotions).

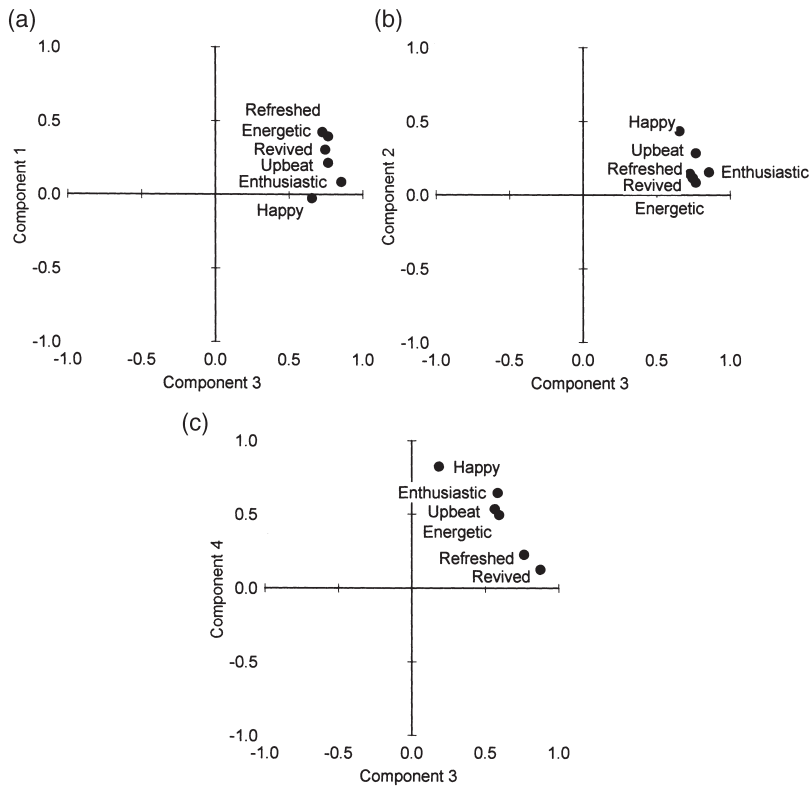

Fig. 2. Pairwise rotated factor plots of the EFI from the three-component solution (panels a and b) and the four-component solution (panel c; components 3, "Positive Engagement", and 4, "Revitalization").

According to this model and other similar formulations (Larsen & Diener, 1992; Tellegen, 1985; Watson & Tellegen, 1985), the adjectives *enthusiastic*, *up-beat*, *energetic*, *refreshed*, and *revived* would fall in the same quadrant of affective space, since they all share the characteristics of high activation and positive valence. For instance, other items considered exemplars of the high activation - pleasantness quadrant in dimensional models include *elated*, *enthusiastic*, *excited*, *peppy*, *strong*, *euphoric*, and *lively* (Larsen & Diener, 1992; Watson & Tellegen, 1985). Therefore, it appears that the distinction between the Positive Engagement and Revitalization components is unwarranted on conceptual grounds.

Second, Gauvin and Rejeski (1993) argued that "feeling energetic or revitalized is not synonymous with nor sufficient for experiencing joy (Shaver, Schwartz, Kirson & O'Connor, 1987)" (p. 411). However, it is important to remember that no evidence was provided that the items in the Positive Engagement and Revitalization components possess content, construct, or discriminant validity as measures of "joy" and "feeling energetic," respectively. Consequently, it is inappropriate to elevate the issue to a conceptual debate when fundamental psychometric questions remain unanswered. Furthermore, the source cited by Gauvin and Rejeski (i.e., Shaver et al., 1987) does not appear to include any supporting evidence for their position. To the contrary, one of the conclusions reached by Shaver and his associates was that "joy is portrayed as energetic, active, and bouncy" (p. 1078).

Third, Gauvin and Rejeski postulated that the lack of a clear distinction between the Positive Engagement and Revitalization items might have been due to the affectively mundane classroom setting where the data for the PCA were collected. To support this argument, they conducted a study involving an acute bout of exercise and analyzed the data using structural equation modeling. The methods and conclusions of this study are examined in the next section.

In summary, the purpose of conducting a PCA at this stage was unclear and the appropriateness of the analytic options is questionable. Most importantly, however, the interpretation of the outcomes of the PCA and, more specifically, the decision to retain a 4-component solution, appears unjustified on psychometric grounds.

Structural equation modeling

Structural equation modeling was used to examine the structure of the EFI following an acute bout of exercise. Specifically, two alternative models were compared: (a) a 3-factor model consistent with the 3-component solution that emerged from the PCA on the basis of the GK rule and (b) a 4-factor model that corresponded to Gauvin and Rejeski's (1993) alleged "a priori conceptual framework for the EFI" (p. 411). In the former case, the items *enthusiastic*, *up-beat*, *happy*, *energetic*, *refreshed*, and *revived* were considered manifestations of a single latent variable, whereas, in the latter case, the former three of these items were considered manifestations of a Positive Engagement latent variable and the latter three were considered manifestations of a Revitalization latent variable. It should be pointed out that no restrictions were placed on either model. Thus, the latent variables, including Positive Engagement and Revitalization, were allowed to correlate freely. In this analysis, the 4-factor model exhibited a better fit. On the basis of a χ^2 difference test (Loehlin, 1998), the improvement in fit was deemed significant and Gauvin and Rejeski concluded that the 4-factor model was "statistically superior" (p. 414) to the 3-factor model and was thus retained as the final structure of the EFI.

There are four problems with this procedure and Gauvin and Rejeski's (1993) line of reasoning in attempting to establish the superiority of the 4-factor model. First, in the framework of scale development, "it …makes no psychometric sense to take a homogeneous pool of substantially intercorrelated items and arbitrarily divide it into separate subscales" (Clark & Watson, 1995, p. 318). Instead, it must be demonstrated that each subscale measures a distinct construct or aspect of a construct, such that intersubscale correlations are substantially lower than intrasubscale correlations. One of the great strengths of structural equation modeling is that "strong theories" can "exploit the effects of being able to fix and/or constrain parameter estimates" (Nunnally & Bernstein, 1994, p. 577). Thus, in testing whether a cluster of items could be meaningfully separated into the Positive Engagement and Revitalization subscales, Gauvin and Rejeski should have fixed or constrained the correlation between the two latent variables to be less than 1.0 (Anderson & Gerbing, 1988; also see Bagozzi, 1993; Church & Burke, 1994, for examples). By not doing so, the indices of fit of the solution were inflated, but the meaningfulness of the solution from a substantive perspective was undermined (Cole, 1987; Floyd & Widaman, 1995). Not surprisingly, the correlation between the Positive Engagement and Revitalization latent variables was very high (i.e., 0.86). Correlations among latent variables are conceived as error-free, so they can be interpreted as having been corrected for attenuation due to unreliability (random measurement error). Similarly, the correlation between the scores on the Positive Engagement and Revitalization scales

was 0.68, which, after a correction for attenuation due to the unreliability of measurement, is raised to 0.90 (Nunnally & Bernstein, 1994). More recently, Lox, Jackson, Tuholski, Wasley and Treasure (2000) reported an uncorrected correlation of 0.857. These values indicate that, when the attenuating effects of random measurement error are taken into account, the constructs assessed by the two subscales are almost perfectly redundant. According to Nunnally and Bernstein (1994), "if a factor correlation is very high … consider replacing the two factors with one… Again, experts differ on 'How high is high?,' but a correlation of .5 should make you consider the option, and .7 would be a very strong reason" (p. 501). Therefore, the distinction between the two subscales appears unjustified.

Second, tests of significance involving a χ^2 difference test should be interpreted with caution. This is because the χ^2 difference test, just like the χ^2 tests used for its calculation (Bentler, 1990; Bollen, 1990; Marsh, Balla & McDonald, 1988), is affected by sample size (Bollen, 1989). Thus, a χ^2 difference test should not be relied upon as the sole reason for choosing one model over another.

Third, Gauvin and Rejeski's (1993) use of the significant χ^2 difference test as a basis for establishing the superiority of the 4-factor solution over the 3-factor solution does not make a compelling case. A model with more parameters will always exhibit a better fit (i.e., lower χ^2) compared to more parsimonious models (Bollen, 1989; Browne & Cudeck, 1992; Hu & Bentler, 1995; Mulaik, et al., 1989). According to Browne and Cudeck (1992), "the difficulty is that the fit of the model can usually be improved by increasing the number of parameters, leading to the temptation to include meaningless parameters that are employed only to give an impression of goodness of fit" (p. 230). Bollen (1989) explained that the value of the fitting function "can be reduced by adding parameters. This is analogous to increasing the R^2 for a regression equation by including more explanatory variables" (p. 270). Therefore, in the case of the EFI, when comparing the 3-factor to the 4-factor solution, one must take into account the reciprocity between goodness of fit and parsimony, as well as that a statistically significant χ^2 difference test does not suffice to establish the "superiority" of one model over another in the absence of substantive conceptual reasons.

Fourth, Gauvin and Rejeski (1993) failed to consider any equivalent models (Breckler, 1990; Cliff, 1983; Cole, 1987; MacCallum, Wagener, Uchino & Fabrigar, 1993). According to MacCallum (1995), "in [structural equation modeling] one is not free to ignore the presence of equivalent models and to assume that one's specified model provides the valid explanation of the data. The problem must be confronted" (p. 31). The notion that a well-fitting "hypothesized" model obliterates the need to examine alternative models was characterized by MacCallum et al. (1993) as "the product of wishful thinking" (p. 197).

To demonstrate the potential pitfalls in Gauvin and Rejeski's (1993) use of structural equation modeling for establishing the "statistical superiority" of the 4-factor solution over the 3-factor solution, we designed a simple computational exercise. First, we used the component loadings from the 4-component PCA solution reported by Gauvin and Rejeski (p. 410) to reproduce the interitem correlation matrix they obtained in their Study 2. Given the high percentage of variance accounted for by this solution (i.e., 79.7%), the reproduced correlation matrix was expected to be very similar to the original. To ensure that our computations were correct, we replicated Gauvin and Rejeski's 4-component PCA. The two solutions were very similar (average discrepancy = 0.03, largest discrepancy = 0.08). Then, we used the reproduced interitem correlation matrix, as

well as the item standard deviations and the sample size from Gauvin and Rejeski's Study 2, to examine some alternative structural equation models. To be consistent with Gauvin and Rejeski's analyses, we also used EQS (Bentler, 1995).

A total of five models were examined. First, we examined Gauvin and Rejeski's 3-factor (Model A) and 4-factor (Model B) models. Then, we examined an alternative 4-factor model (Model C), with the items *enthusiastic*, *upbeat*, *happy*, and *energetic* associated with one factor and the items *refreshed* and *revived* associated with a second factor, presumably driven by common method variance. The factors Tranquility and Physical Exhaustion were retained in the form originally proposed by Gauvin and Rejeski. The purpose of examining this model was to demonstrate that equivalent 4-factor models are plausible. The rationale behind the model was that (a) on the basis of dimensional models (Larsen & Diener, 1992; Russell, 1980; Watson & Tellegen, 1985), the items *enthusiastic*, *up-beat*, *happy*, *energetic*, *refreshed*, and *revived* are essentially homogeneous, but (b) the items *refreshed* and *revived* may form a spurious additional factor driven by common method variance (i.e., semantic overlap, past participle form). This latter suspicion was substantiated by examining the loading patterns from the PCA reported by Gauvin and Rejeski, 1993. Finally, to demonstrate the problems associated with the disregard for the reciprocal relationship between goodness of fit and parsimony, we examined two 5-factor models (Models D and E). Specifically, Model D included separate factors for the items (a) *happy* and *enthusiastic*, (b) *up-beat* and *energetic*, and (c) *refreshed* and *revived*, whereas Model E included separate factors for the items (a) *energetic* and *enthusiastic*, (b) *up-beat* and *happy*, and (c) *refreshed* and *revived*. As with Model C, the factors Tranquility and Physical Exhaustion were retained in the form originally proposed by Gauvin and Rejeski. Furthermore, consistent with Gauvin and Rejeski's analytic approach, we did not impose any additional restrictions on the models and we did not attempt any post hoc modifications.

It is imperative that readers keep in mind that this exercise was based on a reproduced correlation matrix, not on real data, and it was designed for demonstration purposes only. Consequently, it should not be perceived as a substantive test and the findings should not be directly compared to those from Gauvin and Rejeski's original structural equation models. Our results appear in Table 1. The overall fit of the models was unsatisfactory, but this observation is irrelevant for our purposes. Examining the solutions from a comparative standpoint, one can see that Gauvin and Rejeski's finding of improved fit in the 4-factor solution (Model B), compared to the 3-factor solution (Model A), was replicated [χ^2 difference (3, N=256) = 90.87, $P<0$ .001]. As expected,

Table 1
Results from alternative confirmatory factor analyses of the EFI[a]

Model	d.f	χ^2	χ^2/d.f.	NNFI	CFI	GFI	AGFI	SRMSR
A	51	363.24	7.12	0.837	0.874	0.786	0.673	0.062
B	48	272.37	5.67	0.876	0.910	0.834	0.731	0.049
C	48	280.03	5.83	0.872	0.907	0.850	0.756	0.057
D	44	225.46	5.12	0.890	0.927	0.869	0.767	0.050
E	44	223.01	5.07	0.892	0.928	0.872	0.773	0.049

[a] NNFI: Nonnormed goodness of fit index, CFI: comparative fit index, GFI: goodness of fit index, AGFI: adjusted goodness of fit index, SRMSR: standardized root mean square residual. See text for description of models.

however, our alternative 4-factor solution (Model C) also produced an improvement in fit over the 3-factor solution [χ^2 difference (3, N=256) = 83.21, $P<0$ .001], which was almost identical to the one from Gauvin and Rejeski's solution. Also as expected, the introduction of a fifth factor in Models D and E, regardless of its item composition, improved the fit even further, to a statistically significant degree. For both models, the χ^2 difference test showed that the improvement in fit over Gauvin and Rejeski's 4-factor model (Model B), as well as our "alternative" 4-factor model (Model C), was statistically significant ($P<0$ .001).

Although these data should by no means be regarded as substantive findings, they illustrate the problems associated with Gauvin and Rejeski's (1993) reasoning for establishing the supposed "statistical superiority" of the 4-factor solution over the 3-factor solution. In light of these considerations, the conclusion that the factorial validity of the EFI is "strong" (Gauvin & Spence, 1998, p. 333) appears questionable.

Concurrent and discriminant validity

Given the absence of compelling evidence that the Positive Engagement and Revitalization subscales of the EFI can be distinguished on statistical or conceptual grounds, it is not surprising that an examination of the discriminant validity of these scales led to the same conclusion. The correlations of the two scales with the conceptually relevant Positive Affect scale of the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) and the Energy scale of the AD ACL (Thayer, 1989) were very similar.

Furthermore, the Tranquility scale of the EFI was found to have a stronger correlation with the conceptually unrelated Positive Affect scale of the PANAS (an index of high-activation pleasure) than with the conceptually related Calmness scale of the AD ACL (an index of low-activation pleasure). Despite these problems, Gauvin and Rejeski (1993) concluded that "the data from this study, support the position that the subscales of the EFI have good concurrent and discriminant validity" (p. 417).

Recapitulation and conclusions

The various aspects of the conceptual basis of the EFI, as well as the various methodological steps followed in the development of the scale, have been critically reviewed. In both domains, several problems were identified.

From a conceptual standpoint, the following five issues should be considered of primary importance. First, the novel construct of "exercise-induced feeling states" was delineated in terms that were too broad and abstract, thus complicating the process of content and construct validation. Second, given the sketchy present understanding of the nature of affective changes that accompany various types of exercise, the fact that the EFI was developed as a categorical, as opposed to a dimensional, measure was puzzling, particularly since no rationale was presented for this decision. Third, the four constructs that were proposed as the components of "exercise-induced feeling" were not defined and their distinct features were not specified. Furthermore, no rationale was presented for the decision to include only those four states as being relevant to exercise to the exclusion of all other variants of affective experience. Fourth, the notion that exercise is associated

with a universal phenomenology, which cuts across individual and situational differences, is particularly controversial. Combined with the position that this phenomenology can be captured to a meaningfully adequate degree by assessing only four distinct affective states, this notion raises obvious questions of generalizability. The precariousness of this idea becomes evident when one considers the absence of any systematic efforts to investigate the diversity of affective responses to exercise under different conditions. Fifth, the notion of a scale developed to tap only those aspects of affect that are relevant to exercise creates some serious logical problems considering that, in a typical research design, the same scale is also likely to be used in a variety of non-exercise conditions.

The methodological problems that were identified can be seen as a consequence of the aforementioned deficiencies in conceptual groundwork. Although the development of the EFI was initially presented as guided by a theoretical interest in four distinct affective states, the methods that were employed were inconsistent with a deductive approach. At the same time, established guidelines for inductive scale development were not properly followed. Instead, scale development relied heavily on inadequately documented subjective judgements and, most importantly, the content and structure of the EFI were altered by the atypical process of progressively "trimming" the initial item pool. Where traditional psychometric methods were employed, as in the case of the PCA, their purpose was unclear and the analytic options that were used can be challenged, thus creating the potential for serious misinterpretations. Moreover, it appears that the psychometric indices that emerged from these analyses were overlooked in favor of subjective decisions. Finally, problems were identified in Gauvin and Rejeski's (1993) use of structural equation modeling which led to the eventual formulation of the EFI as a 4-factor rather than a 3-factor instrument. Overall, the methods involved in the development of the EFI do not conform to established psychometric standards and certain conclusions appear unjustifiable in light of statistical evidence.

What are the implications of these problems for the utility of the EFI as a measure of affect in exercise research? Given the aforementioned irregularities, we recommend that the EFI not be relied upon as the sole measure of affect. There is no indication that in its present form the EFI can provide a comprehensive assessment of its intended domain of content, namely "exercise-induced feeling." There is also no evidence that the variables measured by the EFI are the "primary forms of affect that are directly influenced by physical activity"(Rejeski et al., 1999, p. 98). It is important to remember that the items considered as relevant to the exercise context by the panel of judges and the sample of students originally formed a total of 15 clusters (assuming that the categorization criteria were valid). Of those 15 clusters describing presumably different affective constructs, only four were eventually selected for inclusion in the EFI, but the criteria for excluding the remaining 11 were not fully disclosed. As a result, it seems that there are numerous "feelings" that people are likely to experience during and following exercise that were not included in the EFI. Most notably, the EFI does not contain scales for assessing negatively valenced states. This consideration is of particular importance when the sample under study does not consist exclusively of young, healthy, and physically active individuals. As the developers of the EFI themselves noted, some clusters of items were dropped because the affective states they represented could be assessed by other psychometric instruments (see Gauvin & Rejeski, 1993, p. 407). For these reasons, reliance on the EFI as the sole measure of affect cannot be recommended. Likewise, conclusions that attempt to generalize findings based on the EFI to the entire domain of "feelings" or affect should be considered inappropriate.

Furthermore, caution should be used when the EFI is employed to assess responses to non-exercise conditions because the irrelevance of some EFI items to such situations may lead to logical and statistical problems. If this practice is unavoidable, researchers are urged to screen their data for violations of distributional assumptions and to cross-examine their findings with alternative measures.

Moreover, there is no apparent conceptual or statistical justification for the distinction between the Positive Engagement and Revitalization items. This is not a problem that is likely to have any serious consequences for the interpretation of the data, but researchers should not expect to derive any useful additional information by scoring these subscales separately.

Given all the aforementioned issues, a reasonable question is whether researchers should consider the EFI as a measurement option. This is for each individual researcher to decide after contemplating the points that we have made and the concerns we have highlighted. In our view, the primary parameter to consider is whether the EFI can provide useful information that cannot be derived from other measures. Considering (a) that "exercise-specificity" is an idea of questionable conceptual merit and practical utility and (b) that the Positive Engagement and Revitalization subscales could be collapsed without loss of useful information, it is interesting to point out that the EFI exhibits a considerable overlap with Thayer's (1989) AD ACL [see Table 4 in Gauvin and Rejeski (1993), p. 416]. In fact, the AD ACL offers the additional advantage of including a Tension subscale, which is of potential relevance, particularly in the context of vigorous exercise.

The development of the EFI represents a pioneering effort. Its publication has helped to rekindle the interest in the study of affective responses to acute exercise and to raise awareness for measurement issues. However, it would be inappropriate to consider the present form of the EFI as a definitive solution without extensive critical analysis. The purpose of the present critique was not to curtail the interest in the measurement of affect, but rather to highlight and clarify some important issues and potential problems. We are convinced that, in the long run, a cautious and critical outlook will prove instrumental in building a secure foundation for progress.

References

American Psychological Association (1985). Standards for educational and psychological testing. Washington, DC: Author.

Anastasi, A. (1990). *Psychological testing.* (6th ed). New York: Macmillan.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Annesi, J. J., & Mazas, J. (1997). Effects of virtual reality-enhanced exercise equipment on adherence and exercise-induced feeling states. *Perceptual and Motor Skills, 85,* 835–844.

Averill, J. R. (1994). I feel, therefore I am - I think. In P. Ekman, & R. J. Davidson, *The nature of emotion: Fundamental questions* (pp. 379–385). New York: Oxford University Press.

Bagozzi, R. P. (1993). An examination of the psychometric properties of measures of negative affect in the PANAS-X scales. *Journal of Personality and Social Psychology, 55,* 836–851.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Bozoian, S., Rejeski, W. J., & McAuley, E. (1994). Self-efficacy influences feeling states associated with acute exercise. *Journal of Sport & Exercise Psychology*, *16*, 326–333.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*, 260–273.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*, 106–148.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*, 214–227.

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, *45*, 16–28.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Newbury Park, CA: Sage.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*, 289–325.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology*, *66*, 93–114.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115–126.

Cliff, N. (1988). The eigenvalue-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*, 276–279.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology*, *55*, 584–594.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology*, *46*, 648–659.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*, 754–761.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis.* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich.

Cronbach, L. J. (1990). *Essentials of psychological testing.* (5th ed). New York: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

DeVellis, R. (1991). *Scale development: Theory and applications.* Newbury Park, CA: Sage.

Ekkekakis, P., Hall, E. E., & Petruzzello, S. J. (1999). Measuring state anxiety in the context of acute exercise using the State Anxiety Inventory: An attempt to resolve the brouhaha. *Journal of Sport & Exercise Psychology*, *21*, 205–229.

Ekkekakis, P., & Petruzzello, S. J. (2000). Analysis of the affect measurement conundrum in exercise psychology: I. Fundamental issues. *Psychology of Sport & Exercise*, *1*, 71–88.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*, 291–314.

Frijda, N. H. (1987). *The emotions.* New York: Cambridge University Press.

Gauvin, L., & Brawley, L. R. (1993). Alternative psychological models and methodologies for the study of exercise and affect. In P. Seraganian, *Exercise psychology: The influence of physical exercise on psychological processes* (pp. 146–171). New York: John Wiley & Sons.

Gauvin, L., & Rejeski, W. J. (1993). The Exercise-Induced Feeling Inventory: Development and initial validation. *Journal of Sport & Exercise Psychology*, *15*, 403–423.

Gauvin, L., Rejeski, W. J., & Norris, J. L. (1996). A naturalistic study of the impact of acute physical activity on feeling states and affect in women. *Health Psychology*, *15*, 391–397.

Gauvin, L., Rejeski, W. J., Norris, J. L., & Lutes, L. (1997). The curse of inactivity: Failure of acute exercise to enhance feeling states in a community sample of sedentary adults. *Journal of Health Psychology*, *2*, 509–523.

Gauvin, L., & Russell, S. J. (1993). Sport-specific and culturally adapted measures in sport and exercise psychology research: Issues and strategies. In R. N. Singer, M. Murphey, & L. K. Tennant, *Handbook of research on sport psychology* (pp. 891–900). New York: Macmillan.

Gauvin, L., & Spence, J. C. (1998). Measurement of exercise-induced changes in feeling states, affect, mood, and emotions. In J. L. Duda, *Advances in sport and exercise psychology measurement* (pp. 325–336). Morgantown, WV: Fitness Information Technology.

Gorsuch, R. L. (1983). *Factor analysis.* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*, 532–560.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*, 149–161.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number of factors rules with simulated data. *Multivariate Behavioral Research*, *17*, 193–219.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*, 238–247.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle, *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Izard, C. E. (1977). *Human emotions.* New York: Plenum.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika*, *35*, 401–415.

Kim, J. O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage.

Kinsman, R. A., & Weiser, P. C. (1976). Subjective symptomatology during work and fatigue. In E. Simonson, & P. C. Weiser, *Psychological aspects and physiological correlates of work and fatigue* (pp. 336–405). Springfield, IL: Charles C. Thomas.

Kline, P. (1994). *An easy guide to factor analysis.* London: Routledge.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark, *Review of personality and social psychology, vol. 13 (Emotion)* (pp. 25–59). Newbury Park, CA: Sage.

Lazarus, R. S. (1991). *Emotion and adaptation.* New York: Oxford University Press.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis.* (3rd ed). Mahwah, NJ: Lawrence Erlbaum.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Lox, C. L., Jackson, S., Tuholski, S. W., Wasley, D., & Treasure, D. C. (2000). Revisiting the measurement of exercise-induced feeling states: The Physical Activity Affect Scale (PAAS). *Measurement in Physical Education and Exercise Science*, *4*, 79–95.

MacCallum, R. (1983). A comparison of factor analysis programs in SPSS, BMDP, and SAS. *Psychometrika*, *48*, 223–231.

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle, *Structural equation modeling: Concepts, issues, and applications* (pp. 16–36). Thousand Oaks, CA: Sage.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*, 185–199.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410.

McAuley, E., Peña, M., Katula, J., & Talbot, H. M. (1997). Affective responses to maximal exercise testing in older adults: Influence of in-task feeling states. *Journal of Sport & Exercise Psychology*, *19*, S83.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual for the Profile of Mood States.* San Diego: Educational and Industrial Testing Service.

Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development*, *30*, 156–164.

Millsar, R. E., & Meredith, W. (1992). Component analysis in multivariate aging research. *Experimental Aging Research*, *18*, 203–212.

Morris, M., & Salmon, P. (1994). Qualitative and quantitative effects of running on mood. *Journal of Sports Medicine and Physical Fitness*, *34*, 284–291.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*, 430–445.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* (3rd ed). New York: McGraw-Hill.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science*, *11*, 341–364.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*, 315–331.

Pedhazur, E. J., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.

Rejeski, W. J. (1994). Dose-response issues from a psychosocial perspective. In C. Bouchard, R. J. Shephard, & T. Stephens, *Physical activity, fitness, and health: International proceedings and consensus statement* (pp. 1040–1055). Champaign, IL: Human Kinetics.

Rejeski, W. J., Gauvin, L., Hobson, M. L., & Norris, J. L. (1995). Effects of baseline responses, in-task feelings, and duration of activity on exercise-induced feeling states in women. *Health Psychology*, *14*, 350–359.

Rejeski, W. J., Reboussin, B. A., Dunn, A. L., King, A. C., & Sallis, J. F. (1999). A modified Exercise-induced Feeling Inventory for chronic training and baseline profiles of participants in the Activity Counseling Trial. *Journal of Health Psychology*, *4*, 97–108.

Rejeski, W. J., & Thompson, A. (1993). Historical and conceptual roots of exercise psychology. In P. Seraganian, *Exercise psychology: The influence of physical exercise on psychological processes* (pp. 3–35). New York: John Wiley & Sons.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*, 403–414.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

Schonemann, P. H. (1990). Facts, fictions, and common sense about factors and components. *Multivariate Behavioral Research*, *25*, 47–51.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*, 1061–1086.

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, *106*, 148–154.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Szabo, A., & Bak, M. (1999). Exercise-induced affect during training and competition in collegiate soccer players. *European Yearbook of Sport Psychology*, *3*, 91–104.

Szabo, A., Mesko, A., Caputo, A., & Gill, E. T. (1998). Examination of exercise-induced feeling states in four modes of exercise. *International Journal of Sport Psychology*, *29*, 376–390.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics.* (3rd ed). New York: Harper Collins.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma, & J. D. Maser, *Anxiety and anxiety disorders* (pp. 681–706). Hillsdale, NJ: Lawrence Erlbaum.

Thayer, R. E. (1989). *The biopsychology of mood and arousal.* New York: Oxford University Press.

Treasure, D. C., & Newbery, D. M. (1998). Relationship between self-efficacy, exercise intensity, and feeling states in a sedentary population during and following an acute bout of exercise. *Journal of Sport & Exercise Psychology*, *20*, 1–11.

Turner, E. E., Rejeski, W. J., & Brawley, L. R. (1997). Psychological benefits of physical activity are influenced by the social environment. *Journal of Sport & Exercise Psychology*, *19*, 119–130.

Veldman, D. J. (1974). Simple structure and the number of factors problem. *Multivariate Behavioral Research*, *9*, 191–200.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1–28.

Vlachopoulos, S., Biddle, S., & Fox, K. (1996). A social-cognitive investigation into the mechanisms of affect generation in children's physical activity. *Journal of Sport & Exercise Psychology*, *18*, 174–193.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*, 219–235.

Widaman, K. F. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*, 354–365.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*, 253–269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442.