Youth Fitness Testing: The Effect of Percentile-Based Evaluative Feedback on Intrinsic Motivation

James R. Whitehead and Charles B. Corbin

This study was a test of Deci and Ryan's (1985) cognitive evaluation theory in a fitness testing situation. More specifically, it was a test of Proposition 2 of that theory, which posits that external events that increase or decrease perceived competence will increase or decrease intrinsic motivation. Seventh and eighth grade schoolchildren (N = 103) volunteered for an experiment that was ostensibly to collect data on a new youth fitness test (the Illinois Agility Run). After two untimed practice runs, a specially adapted version of the Intrinsic Motivation Inventory (IMI) was administered as a pretest of intrinsic motivation. Two weeks later when subjects ran again, they were apparently electronically timed. In reality, the subjects were given bogus feedback. Subjects in a positive feedback condition were told their scores were above the 80th percentile, while those in a negative feedback condition were told their scores were below the 20th percentile. Those in a condition condition received no feedback. The IMI was again administered to the subjects after their runs. Multivariate and subsequent univariate tests were significant for all four subscales dependent variables (perceived interest-enjoyment, competence, effort, and pressure-tension). Positive feedback enhanced all aspects of intrinsic motivation, whereas negative feedback decreased them. In a further test of cognitive evaluation theory, task analysis results supported the prediction that perceived competence would mediate changes in the other IMI subscales. Taken together, these results clearly support cognitive evaluation theory and also may have important implications regarding motivation for those who administer youth fitness tests.

Key words: intrinsic motivation, perceived competence, cognitive evaluation theory, youth fitness testing

Public and professional concerns about youth fitness levels was aroused in the 1950s by the publication of data showing American youth had poor fitness compared to European youth (Kraus & Hirschland, 1954). The consequent media attention led to the formation in 1956 of the President's Council on Physical Fitness (now the President's Council on Physical Fitness and Sports—PCPFS). At the same time the American Association for Health, Physical Education, and Recreation (now the American Alliance for Health, Physical Education, Recreation and Dance—AAHPERD) created a task force to study youth fitness.

The AAHPERD Fitness Task Force conducted a fitness test battery and used it to survey the fitness of American youth and establish national norms. This was achieved for grades 5-12 in 1957 and 1958, and the resultant test—the AAHPERD Youth Fitness Test—became the first physical fitness test with national norms to be developed by the physical education profession in the U.S. (AAHPERD, 1980).

Almost from the start of organized fitness testing, percentile-based award schemes were devised and used (ostensibly) as motivational tools. For example, the AAHPERD YouthFitness Test (which was revised in 1965 and 1976) began giving awards based on the 50th and 80th percentiles in 1965. In 1966 Lyndon Johnson established the President's Council on Physical Fitness (to accompany the AAHPERD test) for those who scored at above the 85th percentile on all the test items.

During a similar time span there were developments in theories of motivation that are pertinent to fitness testing. In 1959, White wrote a seminal paper proposing that people have an intrinsic urge (effectance motivation) toward competence-seeking in the environment. His conceptualization stimulated a wealth of research, the most notable of which was by Harter and his colleagues (e.g., Harter, 1978, 1980; Harter & Connell, 1984) and by Deci and his colleagues (e.g., Deci, 1980; Deci & Ryan, 1985).

Central hypotheses of these developing theories were that changes in the perceived locus of causality of a task behavior and perceived competence at the task influence intrinsic motivation. These tenets were formally stated in Deci and Ryan's (1985) cognitive evaluation theory (pp. 62-64). Recently, the implications of this theory have been noted by exercise scientists and physical educators

Authors' Notes
Both authors contributed equally to the research and consider authorship as joint. The authors wish to extend their appreciation to Eleanor Duddare and Cindy Kauffman, sixth grade physical education teachers at Okemos Middle School, Okemos, MI, for their cooperation in conducting this experiment; to Tom George and Linda Lyman for judging the videotape; and to Kristen Johnson for serving as the peer model. Appreciation is also extended to Daniel Landers and two anonymous reviewers for their comments on an earlier draft of this manuscript.


References

NOTICE: U.S. Copyright Law (Title 17 US code) governs the making of copies of copyrighted material. Provided for Educ. Fair Use.
concerned with designing new fitness education and testing programs. Cognitive evaluation theory suggests the use of extrinsic rewards based on interpersonal (e.g., percent-based) comparisons of fitness may adversely affect intrinsic motivation. This is partly because doing the test to win an award might externalize the locus of causality of the financial reward and, perhaps more importantly, because perceived fitness competence could be reduced by a failure to win the award or by receiving a low percentile ranking on a test. The implication of these theories has provided impetus for the recent construction of three youth fitness education and testing programs. These are Fit Youth Today (American Health and Fitness Foundation, 1986), Fitnessgram (Institute for Aerobics Research, 1987), and Physical best (AAHPERD, 1988). Unlike previous national programs, these three test batteries interpret fitness scores against health criteria and have award schemes based on the reinforcement of positive health-related behavior. To date, the PCFPS has continued to advocate the motivational benefits of percent-based awards and recently has acted independently of AAHPERD for the first time by offering their own Presidential Physical Fitness Award Program (1987).

Although cognitive evaluation theory argues against fitness percent-based comparisons and awards, there does appear to be any direct evidence for or against them in the research literature (Fox & Biddle, 1986, 1987). While there is support for cognitive evaluation and other effectance motivation theories from sport research (e.g., Schubert, 1985; Roberts, Klisnic, & Duda, 1981) and from laboratory studies (Vallerand & Reid, 1984, 1988), this may not apply to youth fitness settings. As Vallerand, Deci, and Ryan (1987) have recommended, research on physical fitness should be conducted separately from research on sports "...since the motivations may be somewhat different in the two settings" (p. 418).

Following that recommendation, the project presented here studied fitness motivation. Its purpose was to examine the effects of intrinsic motivation of receiving different percent-based evaluations of performance after a physical fitness test. More specifically, it was a test of Proposition 2 of Deci and Ryan's (1985) cognitive evaluation theory, which postulates that external events that increase or decrease perceived competence will increase or decrease intrinsic motivation.

Method

Subjects

Seventh and eighth grade students from six separate classes of an urban junior high school in a Southwestern state were invited to take part in this study. Final permission was by parent informed consent. A total of 105 volunteers (72 males, 33 females) participated in the experiment.

Fitness Test

The fitness test used was the Illinois Agility Run. Although strictly an assessment of motor performance (rather than physical fitness), this test was chosen because of its similarity to the commonly used shuttle run test and because it was a novel task for the children. To facilitate the procedures of the experiment, a Lafayette Instrument (Model 65017) reaction/movement timer with a telegraph key starter and photoelectric beam detector finish switch were used for timing.

Instrumentation

An adapted version of the Intrinsic Motivation Inventory (IMI) was used in this study. This instrument (Ryan, 1982) has items specialized written for easy adaptability to a variety of tasks. The original questionnaire had five subscales representing perceptions of interest/enjoyment, perceived competence, effort/importance, pressure/tension, and perceived choice. However, the latter scale was not used since it was experimental—and thus not validated at the time. The time of the test and the identity of the remaining 25 items were chosen for use in this experiment (using language complexity and extant data as selection criteria). Evidence for the validity and adapted versions of the IMI has been demonstrated in sport-related and other experimental situations (e.g., McAvaney, Duncan, & Tammen, 1980, Plant & Ryan, 1985). The instrument utilized a 7-point Likert-type scale. To facilitate interpretation, each of the four subscale scores was calculated by dividing its total score by its number of items.

Procedures

To maintain privacy, subjects left class and came individually to the test site during their regular physical education period. They had been told the purpose of the testing was to obtain normative data for a new youth fitness test. In Phase 1 of the study subjects were allowed two practice runs without timing for the sake of familiarization. The IMI was then administered as a pretest of intrinsic motivation.

In Phase 2 of the study (2 weeks later) the timer appeared to be connected to a portable computer, and the subjects were told that this time the test was "for real." In reality the computer was programmed to give bogus feedback in the form of self-statements (positive, negative, or no feedback) the subjects had been randomly assigned prior to their attempts. As they finished each of their two runs the subjects were shown the "result" on the computer screen. Subjects in the positive feedback condition were given a time that was randomly generated by the computer from within a narrow range typical of high level performance at the test. This was accompanied by the message "Compare to other juniors. Your score is in the top 20% range." Similarly, those in the negative feedback condition were given randomly generated low level scores and the same computer message except that it stated their score was in the "bottom 20% range." These percenttile rankings were verbally reinforced and interpreted as necessary both before and after the two trials. Subjects in the no feedback (neutral) condition saw a message stating "Timing with error," and thus "No scores were recorded." If subjects requested a third attempt to get a time, they were told that because of fatigue the score would not be reliable, but that, if possible, they would be given another attempt during a future session.

Immediately after the second run each subject was again asked to complete the IMI as a posttest of intrinsic motivation. The forms were administered in another room by an assistant who was briefed to prevent subject interaction and to watch for signs of skepticism about the veracity of the feedback. None were observed, and this treatment plausibility was confirmed the following day when a debriefing was conducted. Several subjects even commented on how much they had believed the bogus results.

Design and Statistics

A simple randomized experimental design was used with subjects assigned to one of three conditions: positive, negative, or no feedback groups prior to the Phase 2 testing. This was done separately by gender so that group numbers would be balanced. The eventual slight differences in group numbers were a result of absences from school on the day of testing.

The first task of data analysis was to determine the psychometrics of the revised IMI items. The statistics used for the principal component analysis with varimax rotation, alpha reliability, item-to-scale correlations, and the descriptive statistics. The second task was analysis of treatment effects. Using the data from Phase 2 of the experiment, MANOVA was used to examine the effects of the evaluative feedback on the profile of IMI dependent variables. Subsequently, discriminant analysis and univariate F tests were used to identify specific variables responsible for multivariate differences.

Finally, as a specific test of Proposition 2 of Deci and Ryan (1985) cognitive evaluation theory, path analysis was used to test the prediction that perceived competence would mediate changes in the other aspects (IMI subscales) of intrinsic motivation. The three feedback conditions were dummy coded (Fedhazir, 1982) so that they could be used as predictors in the regression analyses performed to find the path coefficients (beta weights) of the model tested. Specifically, the negative, control, and positive feedback conditions were coded 1, 2, and 3, respectively. The criterion measures in each case were the change scores calculated by subtracting the IMI pretest scores from the posttest scores (see Note 1).

Results

Instrumentation Check

Through a process of item deletion, recalibration, appraisal, further item deletion, and so on, 4 of the originally chosen 20 IMI items were discarded, leaving a total of 16. The main criteria for deletion were improvement in subscale alpha if an item was removed and/or significant cross-factor loadings on the rotated principal components (see Note 2). A principal components analysis of the final 16 items revealed four factors corresponding to the a priori designated subscales, which accounted for 69.9% of the variance (see Table 1). Intercorrelations among the four subscales ranged from .26 to .53 and were all in the expected direction (pressure-tension correlates negatively with the other subscales). Subscale alpha reliability coefficients were from .80 to .88. Finally, the descriptive statistics showed that the sample met satisfactory item means and standard deviations. One of the discarded items was the only "important" item in the originally designated effort/importance scale. Consequently, that amended subscale tapped only perceptions of effort.

Experimental Treatment Effects

A 2 x 3 (Gender x Feedback) MANOVA revealed no gender or gender-treatment interaction effects, so the analysis was collapsed for gender and a one-way MANOVA used to determine the treatment effects on the three groups (see Note 3). A significant multivariate effect for feedback was found, F (8, 196) = 11.47, p < .001. Subsequent univariate F tests (with 2, 102 degrees of freedom) were significant for all four dependent variables (interest/enjoyment: F = 6.36, p < .005, competence: F = 45.30, p < .001, effort: F = 16.15, p < .001, pressure-tension: F = 4.84, p < .01). Examination of cell means showed all changes were in the theoretically predicted direction. Tukey post-hoc tests of significance (p < .05) showed most of these differences were statistically different from one another (see Table 2). Discriminant analysis and further study of the dependent variables within the overall multivariate analysis. Using Wilks's method, perceived competence was the first and only variable entered (F[2, 102]
approximation = 45.50, p < .001). The F values for the remaining three variables were insufficient for them to enter the stepwise analysis. The classification results showed that overall 65.7% of subjects were correctly classified into the three IMI subscales (and their combined total) that resulted from the evaluative feedback, experimental treatments (see Figure 1). In each case the path mediated by perceived competence was statistically and substantively significant (variance explained = 15 to 41%), while the direct path from feedback to the IMI subscale was not (variance explained = 1% or less).

**Perceived Competence as a Mediating Variable**

In each of four fully recursive (unidirectional flow) path analyses, perceived competence mediated the changes in the other three IMI subscales (and their combined total) that resulted from the evaluative feedback, experimental treatments (see Figure 1). In each case the path mediated by perceived competence was statistically and substantively significant (variance explained = 15 to 41%), while the direct path from feedback to the IMI subscale was not (variance explained = 1% or less).

**Discussion**

The purpose of this study was to test the applicability of cognitive evaluation theory to a youth fitness testing situation. For that reason the test performance feedback was purposely interpreted for the subjects in terms of percentile rank classifications as a simulation of the common practice in youth fitness tests.

The results of this experiment supported cognitive evaluation theory. The experimental treatment produced statistically significant changes that were clearly in the theoretically predicted directions. The use of causal modeling techniques particularly supported Proposition 2 of Deci and Ryan’s (1985) theory. The path models revealed changes in perceived competence almost exclusively mediated the changes in the other aspects of intrinsic motivation. From a measurement perspective these findings may also represent an important replication of earlier path analysis studies (Vallerand & Reid, 1984, 1988), which employed the Mayo (1976) Task Reaction Questionnaire (TRQ)—a global measure of intrinsic motivation. A criticism of the TRQ has been that several of its items tap perceptions of competence, its use as a dependent measure (of competence changes) in path models may have merely shown changes in competence on one measure mediating changes in competence in another. The differentiation of aspects of intrinsic motivation into subscales in the IMI allowed removal of this potential confound, and thus the present results can be seen as a validation of the earlier findings.

Interpretation of these results and the implications for implications for youth fitness practitioners requires some discussion of cognitive evaluation theory. Foremost is the theory’s active organism view of human psychological processes. Deci and Ryan (1985) stress the cognitive processing of information (in this case evaluative feedback) through intuitive appraisal and/or reflective judgment. In other words the cognitive evaluation of the information influences the motivational outcomes for an individual.

The fitness test used in this study was a novel task, and thus the subjects had little data on which to base a priori intuitive appraisals of their ability to perform it. It is possible that even if they had formed opinions, the apparent unequivocal precision of the computerized timing might have biased their reflective judgment of self-competence sufficiently to overwhelm their previous expectations. Thus the treatment effect of this experiment may have been particularly powerful due to the somewhat contrived conditions.

However, although the situation may have been somewhat artificial, the effect on perceived competence of evaluative interpersonal comparison and its consequences was relatively clear: Low percentile ranking lowered intrinsic motivation, while high ranking enhanced it. The practical significance of this may be clarified by attention to related theoretical perspectives and also by examination of the ostensible objectives of youth fitness testing.

Since a major stated purpose of fitness testing is to motivate fitness behavior (exercise), it is pertinent to also consider related theories that are concerned with how judgments of personal ability are made during goal-related activities. 

Interpersonal comparisons. Working in the area of interpersonal comparisons for youth fitness practitioners requires some discussion of cognitive evaluation theory. Foremost is the theory’s active organism view of human psychological processes. Deci and Ryan (1985) stress the cognitive processing of information (in this case evaluative feedback) through intuitive appraisal and/or reflective judgment. In other words the cognitive evaluation of the information influences the motivational outcomes for an individual.

The fitness test used in this study was a novel task, and thus the subjects had little data on which to base a priori intuitive appraisals of their ability to perform it. It is possible that even if they had formed opinions, the apparent unequivocal precision of the computerized timing might have biased their reflective judgment of self-competence sufficiently to overwhelm their previous expectations. Thus the treatment effect of this experiment may have been particularly powerful due to the somewhat contrived conditions.

However, although the situation may have been somewhat artificial, the effect on perceived competence of evaluative interpersonal comparison and its consequences was relatively clear: Low percentile ranking lowered intrinsic motivation, while high ranking enhanced it. The practical significance of this may be clarified by attention to related theoretical perspectives and also by examination of the ostensible objectives of youth fitness testing.

Since a major stated purpose of fitness testing is to motivate fitness behavior (exercise), it is pertinent to also consider related theories that are concerned with how judgments of personal ability are made during goal-related activities.

**Table 1. Rotated factor loadings for the IMI items**

<table>
<thead>
<tr>
<th>IMI Subscale</th>
<th>IMI Items</th>
<th>Principal Components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interest-Enjoyment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. I enjoyed doing the fitness test very much</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>5. The fitness test was fun to do</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>9. I thought that the fitness test was a boring activity</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>12. I would describe the fitness test as very interesting</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>17. I thought the fitness test was quite enjoyable</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td><strong>Competence</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. I think I was pretty good at the fitness test</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>3. I think I did pretty well at this test compared to other students</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>10. After doing the fitness test I felt that I was pretty good at it</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td><strong>Effort</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I put a lot of effort into the fitness test</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>7. I didn’t try very hard to do well at the fitness test</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>13. I tried very hard on the fitness test</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>18. I didn’t put much energy into the fitness test</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td><strong>Pressure-Tension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I did not feel nervous at all while doing the fitness test</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>8. I felt very tense while doing the fitness test</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>12. I was very relaxed while doing the fitness test</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>19. I felt pressured while doing the fitness test</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Treatment effects: IMI subscale group means and standard deviations**

<table>
<thead>
<tr>
<th>IMI Subscale</th>
<th>Negative Feedback (n = 37)</th>
<th>No Feedback (n = 37)</th>
<th>Positive Feedback (n = 37)</th>
<th>Overall Mean (n = 101)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interest-Enjoyment</strong></td>
<td>4.44 (1.37)</td>
<td>5.16 (1.09)</td>
<td>5.40 (1.18)</td>
<td>4.98 (1.24)</td>
</tr>
<tr>
<td><strong>Competence</strong></td>
<td>4.30 (1.38)</td>
<td>4.40 (1.09)</td>
<td>4.87 (0.83)</td>
<td>4.69 (1.50)</td>
</tr>
<tr>
<td><strong>Effort</strong></td>
<td>4.81 (1.40)</td>
<td>5.70 (0.83)</td>
<td>6.21 (0.83)</td>
<td>5.58 (1.21)</td>
</tr>
<tr>
<td><strong>Pressure-Tension</strong></td>
<td>3.28 (1.30)</td>
<td>2.81 (1.21)</td>
<td>2.50 (1.11)</td>
<td>2.98 (1.29)</td>
</tr>
</tbody>
</table>

Note. Underlined means (***, ***, *) are statistically similar p < .05.

**Figure 1. Path analysis models showing the role of perceived competence as a mediator of change in other aspects of intrinsic motivation.**

**Perceived Competence**

**Interest-Engagement**

**Perceived Competence**

**Effort**

**Combined Scale**

**Pressure-Tension**

Note. Underlined means (***, ***) are statistically similar p < .05.
comparisons become more salient (Duda, 1987; Nicholls, 1984). This raises a question regarding the extent to which the increases in ego orientation that come with age are an outcome of situational (rather than personal) variables. Classroom-based study (e.g., Deci, Schwartz, Sheinman, & Ryan, 1981) has shown that children taught by control-oriented teachers become lower in intrinsic motivation and self-esteem compared to children taught by autonomy-oriented teachers, and sport-related research has shown that ego-related goals are emphasized more by teachers of older children (Chuamet & Duda, 1988). Since theoretically it should be possible to influence motivational orientation and perceived competence/control through manipulation of these situational variables, further study of them and their influence on motivational outcomes is clearly needed in the exercise/fitness domain.

In summary, although this study supports cognitive evaluation theory, the application of its results to fitness education must be put into perspective in light of the limited amount of other research on youth exercise and fitness motivation. While the results of this experiment support the argument against giving percentile-based interpretations of youth fitness tests (e.g., as in the President’s Challenge), it provides no data that can be used to directly support other procedures such as health-related criterion-based interpretation or recognizing/rewarding exercise behavior. Thus, a main recommendation of this study is that future research investigate the motivational outcomes of those fitness education and testing programs that are based on the exercise mastery process (e.g., Fit youth today, Fitnessgram, Physical Best). In the absence of such research data, it will continue to be difficult to judge the merits of one fitness motivational scheme compared to another.

References


Notes

1. MANOVA was used as a check of the randomization. There were no differences between groups on the pretest scores on the IMI subcales.

2. Sweens’s (1986) recommendation for assessing the statistical significance of factor loadings was followed. This entailed taking the critical value for a simple correlation at an alpha level of .01 (two-tailed and doubling it. This gave an approximate critical value for this analysis of .44. Since a level of .40 is commonly used as a level indicating substantive meaning, it was subjectively decided to critically evaluate any items with cross-loadings above that level. Of the final 16 items, only 1 (item 11) exceeded that figure (.409). The item was retained because its removal would have adversely affected the subscale alpha coefficient.

3. No statistical provision was made for unequal Ns in the gender analysis because proportionality of differences was maintained across cells in the MANOVA.

Authors’ Notes

Thanks are extended to Eddie McCaulley, three anonymous reviewers of initial drafts of this article, and also to Dave Bale, Steve Fedorchak, Kim Meshew, Greg Miller, and Veronica Whitehead for their help with data collection. This study was conducted as part of the first author’s doctoral dissertation under the supervision of the second author.

References


