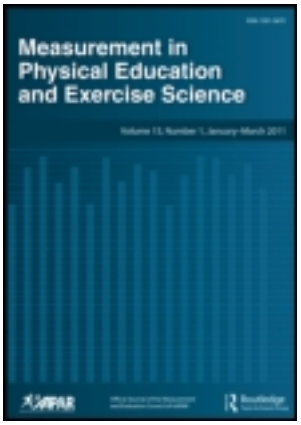

Measurement in Physical Education and Exercise Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmpe20

Tool for Assessing Responsibility-Based Education (TARE): Instrument Development, Content Validity, and Inter-Rater Reliability

Paul M. Wright [a] & Mark W. Craig [b]

[a] Department of Kinesiology and Physical Education, Northern Illinois University, Dekalb, Illinois, USA

[b] Department of Health & Sport Sciences, University of Memphis, Memphis, Tennessee, USA

Published online: 29 Jul 2011.

To cite this article: Paul M. Wright & Mark W. Craig (2011): Tool for Assessing Responsibility-Based Education (TARE): Instrument Development, Content Validity, and Inter-Rater Reliability, Measurement in Physical Education and Exercise Science, 15:3, 204-219

To link to this article: http://dx.doi.org/10.1080/1091367X.2011.590084

Measurement in Physical Education and Exercise Science, 15: 204–219, 2011

ISSN: 1091-367X print / 1532-7841 online
DOI: 10.1080/1091367X.2011.590084

Routledge
Taylor & Francis Group

Tool for Assessing Responsibility-Based Education (TARE): Instrument Development, Content Validity, and Inter-Rater Reliability

Paul M. Wright

Department of Kinesiology and Physical Education, Northern Illinois University, Dekalb, Illinois

Mark W. Craig

Department of Health & Sport Sciences, University of Memphis, Memphis, Tennessee

Numerous scholars have stressed the importance of personal and social responsibility in physical activity settings; however, there is a lack of instrumentation to study the implementation of responsibility-based teaching strategies. The development, content validity, and initial inter-rater reliability testing of the Tool for Assessing Responsibility-Based Education (TARE) are described here. Inter-rater agreement was calculated for paired observations focused on 2 different teachers delivering a total of 18 separate physical education lessons for students in grades 1 through 6. Findings indicate that the Tool for Assessing Responsibility-Based Education provides scores with adequate inter-rater reliability. The procedures employed in this study proved feasible and enable observers to characterize the implementation of responsibility-based teaching in physical education. The Tool for Assessing Responsibility-Based Education has numerous research and training applications relative to the Teaching Personal and Social Responsibility model (Hellison, 2003) and the national content standards for K–12 physical education, specifically Standard 5: Exhibits responsible personal and social behavior that respects self and others in physical activity settings (National Association for Sport and Physical Education, 2004).

Key words: systematic observation, physical education, teaching behaviors, inter-rater agreement

INTRODUCTION

Personally and socially responsible behavior is important for successful learning and development to occur in physical education and other physical activity settings. This assertion is supported by theory, practice, and research (Hellison & Martinek, 2006), as well as the current national content standards for K–12 physical education in the United States that identify "responsible personal and social behavior that respects self and others" as content individuals must learn to become physically educated (National Association for Sport and Physical Education [NASPE], 2004, p. 39). Teaching Personal and Social Responsibility (TPSR; Hellison, 2003) is a well-established instructional model that has been identified as an exemplary approach

Correspondence should be sent to Paul M. Wright, Ph.D., Rm #215 Elma Roane Fieldhouse, Department of Kinesiology and Physical Education, Northern Illinois University, Rm 233 Anderson Hall, Dekalb, IL 60115. E-mail: pwright@niu.edu

to promoting responsibility in physical education and youth sport programs (Metzler, 2005; Petitpas, Cornelius, Van Raalte, & Jones, 2005). Using TPSR as a framework, a number of scholars have demonstrated that responsibility-based instructional strategies integrated systematically into physical activity programs may contribute to a more positive learning environment (Wright & Burton, 2008), higher student motivation (Li, Wright, Rukavina, & Pickering, 2008), more responsible behavior in the gymnasium (DeBusk & Hellison, 1989), as well as transfer of life skills to other settings (Martinek, Schilling, & Johnson, 2001; Walsh, Ozaeta, & Wright, 2010). Despite empirical support and widespread interest in teaching responsibility through physical activity, there is a lack of instrumentation to assess the implementation of responsibility-based teaching strategies in physical education or other physical activity settings (Wright, 2009).

Describing Personal and Social Responsibility in Physical Activity Settings

Hellison's (2003) TPSR model provides a well-developed framework for articulating what constitutes personal and social responsibility in physical activity settings. TPSR is generally described in terms of five responsibility levels or goals: (1) respect for the rights and feelings of others, (2) self-motivation, (3) self-direction, (4) caring, and (5) transfer/outside the gym. The first four levels can be enacted directly in a physical activity program, whereas the fifth level, transfer/outside the gym, relates to transferring the first four levels and associated behaviors to other settings, such as the classroom, playground, or home. Within the TPSR framework, there are numerous examples of specific behaviors associated with each of these. For instance, controlling one's temper, including others, and resolving conflicts peacefully are behaviors that convey respect for the rights and feelings of others. Participating, persisting in challenging tasks, and demonstrating good effort are behaviors that reflect self-motivation. Self-direction is often characterized by working well independently, setting personal goals, and making good decisions. Caring can be displayed through taking on leadership or peer teaching roles and encouraging and helping others. The specific outcomes and behaviors promoted in the TPSR framework align strongly with national physical education content Standard 5: Exhibits responsible personal and social behavior that respects self and others in physical activity settings (NASPE, 2004); however, there is a greater emphasis on transfer outside physical activity settings in TPSR more than in NASPE's Standard 5. In fact, the ultimate goal of the model is teaching youth to take responsibility for the way they conduct themselves and treat others in all aspects of life (Hellison, 2003). While the corresponding national standard acknowledges the importance of promoting responsible behavior outside of physical education, the focus is on transfer to other physical activity settings (NASPE, 2004). Despite this difference, TPSR and national standards overlap considerably in the way they operationalize personal and social responsibility.

Research on Personal and Social Responsibility in Physical Activity Settings

A growing body of research supports the relevance and effectiveness of responsibility-based teaching strategies in physical activity programs. Numerous program evaluations and action research projects have demonstrated the practical effectiveness of TPSR in summer camps, extended day programs, as well as physical education (Hellison & Walsh, 2002). Several peer-reviewed studies have demonstrated TPSR's effectiveness in engaging students and increasing

their capacity to take on responsible roles, including peer-coaching, being in charge of a group, goal setting, and self-evaluation (Cutforth & Puckett, 1999; Walsh, 2008; Wright & Burton, 2008). Although the behavioral norms and values established in a TPSR program may differ from those that students encounter in their home and school culture, TPSR programs appear to be effective in motivating students who have been labeled "at risk" to participate and demonstrate a range of responsible behaviors (DeBusk & Hellison, 1989; Lee & Martinek, 2009; Walsh, 2008; Wright & Burton, 2008). Personal characteristics of the program leader, program features, and responsibility-based teaching strategies appear to foster motivation and a sense of commitment to TPSR programs (Hellison & Wright, 2003; Schilling, 2001; Schilling, Martinek, & Carson, 2007).

The importance of learner responsibility in physical activity settings is also supported by correlational studies. For instance, Li et al. (2008) demonstrated that self-reported levels of personal and social responsibility were positively and significantly correlated with ratings of intrinsic motivation in physical education among middle school students in an urban district. Similarly, Watson, Newton, and Kim (2003) reported that perceptions of responsibility-based values among summer sport camp participants was positively and significantly related to ratings of enjoyment, interest, and positive future expectations in sport. These findings are consistent with those recently reported by Wright and Li (2009) demonstrating that ratings on various scales indicating positive youth development orientation were significantly and positively related to ratings of effort, enjoyment, and belonging in physical education among high school students in an urban district.

Regarding the transfer of responsible behaviors outside of the gymnasium, several studies have demonstrated that discussion, debriefing, and reflection on how it relates to transfer in TPSR programs can advance students' understanding of the life skills being promoted as well as their potential application (Hellison & Wright, 2003; Walsh, 2008; Wright & Burton, 2008). TPSR program evaluations that have incorporated the classroom teachers' perspective indicate that many youth participants display improved behaviors in classroom settings that are at least partially attributed to their involvement in the program (DeBusk & Hellison, 1989; Martinek et al., 2001; Walsh et al., 2010). Although data reported by Martinek and colleagues (2001) were mixed regarding impact on academic achievement, Walsh et al. (2010) reported that classroom teachers consistently indicated that a TPSR extended day program was contributing to participants' academic performance in terms of fewer discipline referrals, better grades, and higher rates of homework completion. In a larger quantitative study involving 122 high school students, Wright, Li, Ding, and Pickering (2010) used a comparison group to assess change on the following variables: tardiness, truancy, grade point average, and disciplinary referrals. Descriptive analysis revealed positive trends on all variables associated with participation in a TPSR program that was integrated into a physical education/health class and delivered on a weekly basis throughout the academic year.

Need for Instrumentation

It is possible that growing empirical support for TPSR will add to the interest in responsibility-based teaching strategies in physical activity settings. Also, the growing emphasis on national standards in the United States should increase the expectation for physical education teachers to articulate exactly how they promote personal and social responsibility (NASPE, 2004). Still, there is a lack of instrumentation to assess the application of responsibility-based teaching

strategies in physical education and other physical activity settings. Wright (2009) suggested that a variety of instruments and methods should be developed to assess TPSR implementation relative to the goals, structure, and processes of the original model (Cummins, Goddard, Formica, Cohen, & Harding, 2003). Meztler (2005) and Rink (2001) have stressed the importance of directly studying the implementation of instructional models and curricular innovations in physical education to test their underlying pedagogical assumptions, to understand their impact on fundamental teaching and learning processes, and to connect those assumptions and processes to learning outcomes.

As noted above, the national standards mandate that K–12 physical education teachers promote personally and socially responsible behavior (NASPE, 2004). Efforts are underway to develop instrumentation related to all of the national standards, but the least progress has been made regarding standards that address the affective learning domain (see Erwin & Castelli, 2008). It is difficult to assess students in these areas when many teachers are unsure how to teach to these standards in systematic and purposeful ways (Parker & Hellison, 2001; Parker, Kallusky, & Hellison, 1999; Parker & Steihl, 2005). This underscores the need for tools that assess the implementation of responsibility-based teaching strategies. In addition to research and evaluation, such tools could also have applications for teacher training and professional development. Therefore, the purpose of this research was to develop and assess the content validity and inter-rater reliability of an instrument that enables observers to characterize the implementation of responsibility-based teaching in physical education and other physical activity settings.

METHODS

Instrument Development

The first step in developing the Tool for Assessing Responsibility-Based Education (TARE) was determining the instrument's content. The first author drew from over ten years of experience as a TPSR practitioner/researcher, as well as the extant literature and NASPE's (2004) descriptions of personal and social responsibility in physical education, to draft the core content of the TARE. The second step was to generate items that would align this content with a systematic observation methodology. For example, instructional strategies often used to promote responsibility were operationalized as discrete observable teaching behaviors. This process of item generation was informed by a review of several well-established systematic observation instruments. Valid and reliable instruments focusing on student and/or teacher behaviors in physical education included the Academic Learning Time-Physical Education (ALT-PE; Parker, 1989) and the System for Observing Fitness Instruction Time (SOFIT; McKenzie, Sallis, & Nader, 1991). Model instruments validated in classroom and school-wide applications included the Classroom Observation Measure (COM) and the School Observation Measure (SOM) developed at the Center for Research in Educational Policy (Lewis, Ross, & Alberg, 1999; Ross, Smith, Alberg, & Lowther, 2004; Sterbinsky & Ross, 2003).

Pilot Study and Field Testing

After a draft instrument was prepared, the procedures for data collection and coding were developed using time-sampling methods that have been shown to be effective in the previously cited

instruments as well as the System for Observing Play and Leisure Activity in Youth (SOPLAY; McKenzie, Marshall, Sallis, & Conway, 2000) and the System for Observing Play and Recreation in Communities (SOPARC; McKenzie, Cohen, Sehgal, Williamson, & Golinelli, 2006). The draft instrument and procedures were pilot tested by the authors using video footage of TPSR lessons delivered by the first author. This led to the refinement of the original content and operational definitions. This process also provided preliminary validation of the TARE's content in that all indicators of effective TPSR implementation were observed in lessons delivered by an expert practitioner. At this point, the content and procedures appeared feasible and ready for field testing.

The instrument was field tested in four secondary physical education classes. Live observations were conducted in classes delivered by three different physical education teachers in the same urban public high school. All the classes observed were co-educational and made up of students from racially and economically diverse backgrounds. Generally, there were between 25 and 35 students in each class. None of teachers were formally implementing the TPSR model but did represent a range of effectiveness in promoting personally and socially responsible behavior. The authors served as observers and openly discussed what they were seeing and how determinations should be made regarding the TARE ratings while observing the first two classes, each taught by a different teacher. At that point, it became clear that the observers had reached a shared understanding of the operational definitions, coding, and procedures. In observing the third and fourth classes, both taught by the remaining teacher, the observers did not discuss or share their ratings during the process. The independent ratings from these last two observations exceeded 80% inter-rater agreement, the standard promoted by Krippendorf (1980).

Content Validation

Field testing supported the TARE's content validity, as the more empowerment-based teaching strategies were rarely observed, if ever. This indicated that the TARE could discriminate between a robust implementation of responsibility-based pedagogy, such as that seen in video-taped TPSR lessons compared to typical physical education instruction. At this point, the content of the instrument was finalized, and the procedures had proven feasible in the field. Thus, the TARE was deemed ready for more formal content validity and inter-rater reliability testing.

To formally assess content validity, the TARE was presented to a panel of experts for review. The panel included the TPSR model developer, two established physical education pedagogy experts, one former physical education administrator from a large urban school district, and one community health researcher with experience assessing implementation of community-based physical activity interventions. This panel presented a range of experience and was highly qualified to assess the content validity of the instrument relative to the TPSR model as well as the related national standards. Three of the panelists had experience with other systematic observation tools such as the ALT-PE, SOFIT, and the SOM. The first author had face-to-face meetings with all members of the panel over a period of several months. In each case, the first author described the purpose and development of the TARE, provided an overview of the content and described the data collection procedures. Following that introduction, panel members were asked to review the TARE's content in detail, ask for any clarification they required, and provide feedback regarding any errors, omissions, or ambiguity. Panel members were also invited to comment on the rigor and feasibility of data-collection procedures. All panelists provided positive reviews of the content and data collection procedures in these meetings. One of the pedagogy experts

provided useful feedback to make certain operational definitions more explicit. These changes in wording were made to the panelist's satisfaction without altering the original meaning. All panelists were invited to contact the first author if, upon further review, they had additional questions or suggestions; none did.

Structure of Final Instrument

The first major section, Observable Teaching Strategies, is an interval recording system that requires observers to make rating decisions based on what the teacher does and says. The observers pay strict attention to the teacher and determine whether or not they see evidence related to nine discrete teaching strategies during a 5-min period. While time sampling tools focused on physical activity levels often use much shorter intervals (less than 1 min) and the SOM uses a 15-min period, video analysis and field testing indicated that a 5-min time interval was large enough yet sufficiently sensitive to rate the usage of the various teaching strategies used in the TARE (McKenzie et al., 1991; Sterbinsky & Ross, 2003). These strategies (see Table 1) represent a range of teacher behaviors that promote or foster personal and social responsibility. Some of these strategies are fundamental to good teaching, and others involve a greater degree of student responsibility than is typically seen in K–12 physical education (Doolittle & Demas, 2001; Hellison, 2003; Hellison, Cutforth, Martinek, Kallusky, Parker, & Steihl, 2000; Parker & Hellison, 2001; Parker et al., 1999). After each 5-min interval, the appropriate codes on a scoring sheet are circled to indicate which strategies were observed during that interval (see Table 2). Some strategies, such as modeling respectful behavior, may be employed throughout the interval, while other strategies, such as assigning a specific task to a student, may be displayed in a single discrete action. In either case, the code is circled once to indicate that the strategy was employed. All strategies that occur in a given 5-min interval are coded.

Section Two

Section Two, Personal–Social Responsibility Themes, is completed after the last 5-min interval for a given lesson has been coded. In this section, observers provide a holistic assessment of the extent to which the teacher promoted responsibility throughout the lesson. Ratings are made relative to four themes that characterize teaching for personal and social responsibility (Hellison, 2003). These themes are integration, transfer, empowerment, and teacher–student relationship. Ratings are based on a 5-point Likert scale ranging from 0 (*Never*) to 4 (*Extensively*). The scoring sheet for this section, including theme definitions, can be seen in Table 3.

Section Three

Section Three, Student Responsibility, is also completed after the last 5-min interval in a lesson has been coded. This section requires observers to assess the degree to which students displayed personally and socially responsible behavior during the lesson. The criteria used in this section relate directly to the student behaviors called for in the TPSR model (Hellison, 2003), as well as the corresponding national standard (NASPE, 2004). These behaviors, their definitions and the 5-point rating scale used to assess them can be seen in Table 4.

TABLE 1
Extended Description of Responsibility-Based Teaching Strategies

Modeling respect (M): Teacher models respectful communication. This would involve communication with the whole group and individual students. Examples include using students' names, active listening, making eye contact, recognizing individuality, maintaining composure, developmentally appropriate instruction, talking 'with' rather than 'at' students, showing an interest in students, and unconditional positive regard. Counter-examples include indifference, disengagement, losing temper, and deliberately embarrassing a student.

Setting expectations (E): Teacher explains or refers to explicit behavioral expectations. Examples include making sure all students know where they should be and what they should be doing at any given time; giving explicit expectations for activity or performance; explaining and reinforcing safe practices, rules, and procedures, or etiquette.

Opportunities for success (S): Teacher structures lesson so that all students have the opportunity to successfully participate and be included regardless of individual differences. Examples in physical activity include making appropriated adaptations for inclusion and providing opportunities for practice, skill refinement, and game play. Examples in less active modes include allowing students to volunteer answers in a discussion or succeed in a non-physical task.

Fostering social interaction (SI): Teacher structures activities that foster positive social interaction. Examples include fostering student–student interaction through cooperation, teamwork, problem solving, peer-coaching, partner drills where communication is encouraged, and conflict resolution or debriefing. Counter-examples include random student interactions not fostered or supported by the teacher and pseudo group discussions that only involve student–teacher exchanges.

Assigning management tasks (T): Teacher assigns specific responsibilities or management-related tasks that facilitate the organization of the program or a specific activity. Examples include asking students to take attendance, serve as timekeeper, set up equipment, keep score/records, or officiate a game.

Leadership (L): Teacher allows students to lead or be in charge of a group. Examples include allowing students to demonstrate for the class, lead a station, teach/lead exercises for the whole class, or coach a team.

Giving choices and voices (V): Teacher gives students a voice in the program. Examples include letting students engage in group discussions, vote as a group, and make individual choices; inviting student questions or suggestions, eliciting student opinions, and letting students evaluate the teacher or program.

Role in assessment (A): Teacher allows students to have a role in learner assessment. Examples include self- or peer-assessment related to skill development, behavior, attitude, etc.; student-centered goal-setting; and negotiation between teacher and student on their grade or progress in the class.

Transfer (Tr): Teacher directly addresses the transfer of life skills or responsibilities from the lesson beyond the program. Examples of topics include the need to work hard and persevere in school; the importance of being a leader in your community; keeping self-control to avoid a fight after school; setting goals to achieve what students want in sports or life in general; the need to be a good team player when in other contexts, such as the workplace; the value of thinking for yourself to avoid peer pressure and make good life choices.

Reliability Testing

Setting and Participants

The setting for reliability testing was an elementary school serving grades 1–6 in a mid-sized southern city. Although officially part of the public school district, this is a laboratory school affiliated with a large university. Each class section in the school has two 30-min physical

TABLE 2
Scoring Sheet for Section One: Observable Teaching Strategies

Time Intervals	*Responsibility – Based Strategies*	*Comments*
0–5	M E S SI T L V A Tr	
5–10	M E S SI T L V A Tr	
10–15	M E S SI T L V A Tr	
15–20	M E S SI T L V A Tr	
20–25	M E S SI T L V A Tr	
25–30	M E S SI T L V A Tr	
30–35	M E S SI T L V A Tr	
35–40	M E S SI T L V A Tr	
40–45	M E S SI T L V A Tr	
45–50	M E S SI T L V A Tr	
50–55	M E S SI T L V A Tr	
55–60	M E S SI T L V A Tr	
60–65	M E S SI T L V A Tr	
65–70	M E S SI T L V A Tr	
70–75	M E S SI T L V A Tr	
75–80	M E S SI T L V A Tr	
80–85	M E S SI T L V A Tr	
85–90	M E S SI T L V A Tr	

Note: Codes: M—modeling respect, E—setting expectations, S—opportunities for success, SI—fostering social interaction, T—assigning management tasks, L—leadership, V—giving choices and voices, A—role in assessment, Tr—transfer.

education lessons weekly. Reliability testing for the TARE focused on 18 different lessons taught by 2 different instructors, 1 an expert and the other a novice. The expert teacher had 26 years of teaching experience, held a master's degree, and had recently been honored as the state's elementary physical education teacher of the year. The novice was completing a semester of student teaching as the culminating experience in an accredited K–12 physical education licensure program.

A total of eight class sections, representing grades 1–6, were involved in this study during the spring of 2008. Several of these class sections were observed more than once. Class sizes varied but were generally between 18 and 22. A total of 215 different students (63% white and 37% non-white/minority; 52% male and 48% female) were involved in the observations. Some students with mild or moderate disabilities were involved and were fully included. This study was approved by the university's Institutional Review Board and the school's administration.

Procedures

The authors are the same trained observers who had reached 80% inter-rater agreement in the pilot study. Blinded paired observations of 93 5-min intervals were conducted across 18 lessons to formally assess reliability in terms of inter-rater agreement. Observers sat several feet away from each other but close enough to have a similar vantage point and cue each other regarding the

TABLE 3
Scoring Sheet for Section Two: Personal–Social Responsibility Themes

	4—Extensively	*3—Frequently*	*2—Occasionally*	*1—Rarely*	*0—Never*	*Comments*
Integration: extent to which responsibility roles and concepts are integrated into the physical activity	4	3	2	1	0	
Transfer: extent to which connections are being made to the application of life skills in other settings	4	3	2	1	0	
Empowerment: extent to which the teacher shares responsibility with students	4	3	2	1	0	
Teacher–student relationship: extent to which students are treated as individuals deserving respect, choice, and voice	4	3	2	1	0	

Note: Extensively—theme is seamlessly addressed directly and evidenced in multiple ways throughout the lesson through the words and actions of the teacher. *Frequently*—theme is addressed directly and evidenced at several points in the lesson through the words and actions of the teacher. *Occasionally*—some of the teacher's words and actions connect to this theme either directly or indirectly during the lesson. *Rarely*—this theme is not generally integrated into the teaching but may be reflected in some isolated words or actions on the teacher's part. *Never*—throughout the entire lesson, none of the teacher's words or actions clearly convey or align with this theme.

start- and end-points of the 5-min intervals. As a class was entering the gymnasium and waiting for the lesson to begin, the date, grade level, schedule, teacher characteristics, number of students, and other important contextual information were recorded on a data sheet. When the bell rang to signal the beginning of the class period, the first 5-min interval was begun using a stopwatch. In each interval, the primary focus was on the teacher's words and actions, where evidence was sought of the nine teaching strategies represented in Section One of the TARE. The coding of 5-min intervals continued until the lesson was clearly finished and students began lining up to prepare for dismissal and transition to their next destination. If the last observed interval was less than 3 min, it was discounted from the dataset. As the lesson was being observed, contextual notes describing the lesson content and activities were made in a space for open comments on the scoring sheets. Also noted were specific examples of how strategies were implemented. For instance, if *Leadership* was coded in a particular 5-min interval, a note was made to support that decision (i.e., "student was designated team captain"). After the last interval was coded, the raters completed Sections Two and Three based on their overall impression of the lesson. Frequently, between lessons, raters debriefed on what they had observed and how their coding decisions did or did not agree. This process allowed for ongoing calibration and prevented deviations or observer drift. When it became obvious that characteristic patterns were emerging for the teachers, observers debriefed and reviewed definitions to ensure they were basing decisions on operational definitions rather than habit or assumption.

TABLE 4
Scoring Sheet for Section Three: Student Responsibility

	4—Very Strong	*3—Strong*	*2—Moderate*	*1—Weak*	*0—Very Weak*	*Comments*
Respect: student does no harm to others verbally or physically, includes/works well with others, resolves conflicts peacefully if they emerge	4	3	2	1	0	
Participation: student will try every activity and take on various roles if asked	4	3	2	1	0	
Effort: student tries hard to master every task and focuses on improvement	4	3	2	1	0	
Self-direction: student will stay on task without direct instruction or supervision whether working alone or with others, does not seem to follow bad examples or peer pressure	4	3	2	1	0	
Caring: student will help, encourage others, and offer positive feedback	4	3	2	1	0	

Note: Very strong—all students displayed this responsibility throughout the lesson with no observed exceptions. *Strong*—most students displayed this responsibility throughout the lesson with only minor and/or isolated exceptions. *Moderate*—many students displayed this responsibility, but many did not; several exceptions were observed. *Weak*—some students displayed this responsibility, but many did not; exceptions were frequent and/or serious enough to impede learning. *Very weak*—few, if any, students displayed this responsibility while the majority struggled to do so; exceptions were frequent and/or serious enough that at least some portions of the lesson were rendered ineffective.

Data Analysis

Following the example of several well-established observation tools that use interval ratings, the TARE's inter-rater reliability was assessed by calculating the percent agreement between independent observations. There are varying opinions regarding benchmarks for an acceptable level of reliability using this approach (Neuendorf, 2002). Krippendorf (1980) proposed that reliability reporting for any reliability coefficient should be done only if the reliability is above .80 with tentative reporting if the variable is between .67 and .80. Riffe, Lacy, and Fico (1998) also endorsed a high standard, stating that researchers typically report reliabilities in the .80 or higher range and that those variables below .70 are hard to interpret and duplicate. For assessing the reliability of the interval ratings from Section One of the TARE, the latter guidelines were used. The Kappa statistic was considered as an option for analyzing the interval rating data. This statistic addresses the critique that percent agreement alone is not a sufficient measure for inter-rater reliability, as it does not factor chance into the calculations (Cohen, 1968); however, application of the Kappa statistic proved inappropriate in this study for two reasons. First, the Kappa statistic was designed for situations where multiple codes are applied as opposed to a

binary (observed versus not observed) coding system like that used in Section One of the TARE. Second, and more importantly, the Kappa statistic was not usable for some variables in this data set because complete (100%) agreement was achieved for some indicators, and this situation renders the Kappa statistic uncalculable.

Following Lewis et al. (1999), reliability on the holistic items in Sections Two and Three was calculated as percent of ratings that were in agreement or differed by only one point on the 5-point rating scale. Ratings in these sections are more general because they assess multiple behaviors and interactions that may be observed throughout the entire lesson; therefore, it is not reasonable to expect the same level of precision seen in the interval ratings. The contribution of ratings in Sections Two and Three is contextual information. As this information is quite valuable, it was deemed important to assess the consistency of the data, even if using a less stringent standard. This approach is supported by Uebersax (1992), who argued that when testing inter-rater reliability on a Likert scale, the main purpose is to assess whether or not the raters are consistent and not necessarily if they are in exact agreement.

RESULTS

Results are presented for the novice and expert teacher, separately as well as the total, from all observations. Table 5 displays the percent agreement achieved on the observable teaching strategies from Section One. These calculations are based on a total of 93 intervals that were observed and rated across 18 lessons. On the nine strategies in this section, the percent agreement was consistently high. Results from the 43 intervals in which the novice was observed ranged from 93.0% to 100.0%, and results from the 50 intervals in which the expert was observed ranged from 80.4% to 100.0%. When aggregated, data from the 93 intervals yielded percent agreements ranging from 88.3% to 100.0%. The reliability results from the novice and expert were similar overall. The differences between novice and expert results were larger relative to *setting expectations* and *giving choices and voices*. Accordingly, these are the two indicators with the lowest level of total agreement. Still, all ratings in Table 5, whether for the novice, expert, or aggregate, exceeded the .80 benchmark for reliability called for by Krippendorf (1980).

As explained earlier, the holistic ratings made at the end of each observed lesson were assessed for consistency using a less stringent standard. Instead of exact agreement, the extent to which the observers provided ratings within one point on the 5-point scale was calculated. Table 6 displays the consistency of ratings for the personal–social responsibility themes from Section Two. The only instances for individual teachers when results were below 90.0% agreement within one point were the ratings of empowerment and student–teacher relationship for the novice; however, these results (75.0% and 87.5% agreement within one point, respectively) were still largely consistent. When aggregated, the results across 18 lessons for Section Two items ranged from 88.9% to 100.0% agreement within one point. Table 7 displays the results for ratings of student responsibility from Section Three. In all instances except one, results were 87.5% agreement within one point or higher on the individual teacher ratings. The exception was ratings of self-direction for the expert teacher, which was only 70% agreement within one point . Although these ratings emerged as relatively low, the aggregated data show ratings on this item were 77.8% agreement within one point . In sum, across 18 lessons, the items in Section Three ranged from 77.8% to 100.0% agreement within one point.

TABLE 5
Percent Agreement for Observable Teaching Strategies

	Percent Agreement		
	Novice (43 Intervals)	*Expert (50 Intervals)*	*Total (93 Intervals)*
Modeling respect	100.0%	98.0%	98.9%
Setting expectations	97.7%	80.4%	88.3%
Opportunities for success	95.3%	100.0%	97.9%
Social interaction	93.0%	96.1%	94.7%
Assigning management tasks	95.3%	100.0%	97.9%
Leadership	100.0%	100.0%	100.0%
Choices and voices	100.0%	84.3%	91.5%
Role in assessment	100.0%	100.0%	100.0%
Transfer	100.0%	100.0%	100.0%

TABLE 6
Consistency for Personal–Social Responsibility Themes

	Percent of Ratings within One Point on a 5-Point Scale		
	Novice (8 Lessons)	*Expert (10 Lessons)*	*Total (18 Lessons)*
Integration	100.0%	90.0%	94.4%
Transfer	100.0%	100.0%	100.0%
Empowerment	75.0%	100.0%	88.9%
Relationship	87.5%	100.0%	94.4%

TABLE 7
Consistency for Student Responsibility Levels

	Percent of Ratings within One Point on a 5-Point Scale		
	Novice (8 Lessons)	*Expert (10 Lessons)*	*Total (18 Lessons)*
Respect	100.0%	90.0%	94.4%
Participation	100.0%	90.0%	94.4%
Effort	87.5%	90.0%	88.9%
Self-direction	87.5%	70.0%	77.8%
Caring	100.0%	100.0%	100.0%

DISCUSSION

The purpose of this research was to develop and evaluate the content validity and inter-rater reliability of a comprehensive instrument that would allow observers to characterize the implementation of responsibility-based teaching in physical education and other physical activity

settings. The primary component of the TARE, Section One, makes use of a time interval sampling methodology to document the use of specific responsibility-based teaching strategies. Results reported here provide evidence of the content validity and the inter-rater reliability of the scores derived from this section of the TARE in the population studied. The percent agreement on all nine items in this section exceeded the stringent .80, benchmark for reliability called for by Krippendorf (1980). For two of the items, the level of agreement was notably lower for the expert. This may relate to the fact that these observations were conducted earlier in the data collection process when a shared understanding of the operational definitions was still being developed. Nonetheless, the overall level of agreement was strong on all items in this category and comparable to well-established instruments designed to assess teacher and student behavior in physical education (McKenzie et al., 1991; Parker, 1989).

Sections Two and Three call for holistic ratings to be made at the end of an observed lesson. The former calls for global ratings of the teacher's implementation of responsibility-based teaching, and the latter assesses the level of the students' responsibility in the lesson. The Likert scale ratings used in these sections are important for providing context but are unlikely to yield the same high level of inter-rater reliability as that seen in the time interval sampling section (Uebersax, 1992). Therefore, consistency was assessed in these sections using the guideline of ratings being within one point of each other on a 5-point scale. Using this standard, results in both of the holistic rating sections proved reasonably consistent. The least consistent ratings related to self-direction in the student responsibility section. This may reflect the difficulty of developing a precise operational definition for an aspect of responsibility that could manifest in many different ways. Even this item was rated with 77.8% agreement within one point overall. It should also be noted that data in the current study were analyzed and reported at the item level. If the data are aggregated by section, it is apparent that the consistency of the holistic rating sections of the TARE are comparable to the SOM, a well-developed tool for assessing teacher behaviors in the classroom setting (Ross et al., 2004). Lewis et al. (1999) assessed inter-rater agreement on all 27 SOM items combined. Each item represented a specific teaching behavior associated with best practice such as making use of cooperative learning strategies. They reported perfect agreement in 66.7% of the paired ratings, agreement within one point in 93.7% of the paired ratings, and agreement within two points in 100% of the paired ratings. If results from Section Two of the TARE are combined in this fashion, the overall level of agreement is 94.4% (68/72) agreement within one point across the four items. Calculated the same way, the overall level of agreement for Section Three of the TARE is 91.1% (82/90) agreement within one point across the five items.

Heretofore, studies related to TPSR have been inconsistent in the way they address implementation of the curricular model. Wright (2009) suggested that a more explicit understanding of implementation is important for this growing line of research. TPSR studies that have addressed implementation directly have employed varying combinations of teacher self-report, documentation, ethnographic interviews, and open field notes (Buchanan, 2001; Walsh et al., 2010; Wright & Burton, 2008; Wright et al., 2010). To complement these methods and add consistency to the framing of implementation in TPSR studies, it is recommended that TPSR researchers integrate the TARE into their research designs. This instrument need not supplant other methods but, if integrated, could enhance the researchers' ability to describe and assess the implementation of a TPSR program or intervention.

While the TARE was framed largely around the TPSR model, its application is not restricted. The instrument was designed to systematically obtain observational data that can reflect the extent to which responsibility-based teaching strategies are being employed and enacted in a physical activity lesson. Therefore, the TARE is directly relevant to K–12 physical education content standards in the United States (NASPE, 2004). While research on these national standards is growing, no systematic research has been conducted to date on Standard 5 (exhibits responsible personal and social behavior that respects self and others in physical activity settings). The current study demonstrates the feasibility of using the TARE in physical education settings. This instrument is recommended for use in studies that address the full spectrum of content in physical education, including personally and socially responsible behavior. The TARE could be used to assess teachers' implementation of the corresponding standard. Student performance in this area could be assessed, at least in part by a recently published instrument that demonstrated that it produced reliable and valid scores of self-reported personal and social responsibility in physical education (Li et al., 2008), as well as a number of authentic assessments generated by TPSR scholars (Hellison, 2003).

In addition to research and evaluation activities, it is proposed that the TARE could be useful as a training tool. Faculty in physical education teacher education programs could use this instrument in training pre-service teachers to implement the TPSR model or simply to enact the full spectrum of physical education content mandated in the national standards. Possible applications include having pre-service teachers use the TARE to conduct video assisted self-assessments or real-time peer-assessments. This approach is commonly used with instruments such as the ALT-PE to increase pre-service teachers' awareness of their use of time. It is posited that in a comparable way, the TARE could help pre-service teachers become more aware of how they can be proactive and purposeful in promoting personal and social responsibility.

Findings presented here indicate the the TARE merits further development in terms of reliability testing and instrument validation. Although the levels of inter-rater reliability for Sections Two and Three were lower, this was deemed acceptable, as data from these sections are contextual in nature. The current study is limited by a relatively small number of observations and the fact that the observers were also the instrument developers and the researchers. Future researchers should assess inter-rater reliability based on a larger number of observations conducted by multiple pairs of trained observers with no potentially conflicting roles. To achieve this, and to facilitate greater use of the instrument, a training manual should be developed for using the instrument. It should also be noted that only inter-rater reliability was assessed. If lessons had been video-taped, it would have been possible to also assess intra-rater reliability. This should also be considered by future researchers. As a means of supporting implementation of responsibility-based teaching strategies in practice, a post-teaching reflection tool reflecting the same content should be developed. Not only would such a tool be useful to practitioners and pre-service teachers, but it could provide an additional data source for triangulation in research studies that include the TARE observation tool in the design. Finally, while field testing during the instrument development stage reflected well on the TARE's usefulness at the secondary level and with larger class sizes, future studies should formally validate the instrument with different grade levels and varying class sizes.

In summary, data presented here indicate that the TARE is a comprehensive instrument that allows observers to characterize the implementation of responsibility-based teaching in physical education and other physical activity settings. Several levels of support have been provided for

the content validity of the instrument relative to the TPSR model and the related national content standard for physical education. Inter-rater reliability for the individual items in the time interval sampling portion of the instrument meet a rigorous standard and are comparable to that reported for similar instruments used in physical education and physical activity settings. Items contained in the sections of the TARE that call for holistic ratings also appear to be consistent, albeit by a less rigorous standard. The TARE has numerous research and training applications relative to the TPSR model and K–12 physical education.

REFERENCES

Buchanan, A. M. (2001). Contextual challenges to teaching responsibility in a sports camp. *Journal of Teaching in Physical Education*, *20*, 155–171.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–219.

Cummins, M., Goddard, C., Formica, S., Cohen, D., & Harding, W. (2003). *Assessing program fidelity and adaptations toolkit* (pp. 3–9). Newton, MA: Health and Human Development Programs, Educational Development Center, Inc.

Cutforth, N., & Puckett, K. (1999). An investigation into the organization, challenges, and impact of an urban apprentice teacher program. *The Urban Review*, *31*, 153–172.

DeBusk, M., & Hellison, D. (1989). Implementing a physical education self responsibility model for delinquency prone youth. *Journal of Teaching in Physical Education*, *8*, 104–112.

Doolittle, S., & Demas, K. (2001). Fostering respect through physical activity. *Journal of Physical Education, Recreation, and Dance*, *72*, 37–41.

Erwin, H. E., & Castelli, D. M. (2008). National physical education standards: A summary of student performance and its correlates. *Research Quarterly for Exercise and Sport*, *79*, 495–505.

Hellison, D. (2003). *Teaching responsibility through physical activity* (2nd ed.) (pp. 15–38). Champaign, IL: Human Kinetics.

Hellison, D., Cutforth, N., Martinek, T., Kallusky, J., Parker, M., & Steihl, J. (2000). *Youth development and physical activity: Linking universities and communities* (pp. 31–49). Champaign, IL: Human Kinetics.

Hellison, D., & Martinek, T. (2006). Social and individual responsibility programs. In D. Kirk, D. Macdonald, & M. O'Sullivan (Eds.), *The handbook of physical education* (pp. 610–626). Thousand Oaks, CA: Sage.

Hellison, D., & Walsh, D. (2002). Responsibility-based youth programs evaluation: Investigating the investigations. *Quest*, *54*, 292–307.

Hellison, D., & Wright, P. M. (2003) Retention in an urban extended day program: A process-based assessment. *Journal of Teaching in Physical Education*, *22*, 369–381.

Krippendorf, K. (1980). *Content analysis: An introduction to its methodology* (2nd ed., Chapter 11). Beverly Hills, CA: Sage.

Lee, O., & Martinek, T. (2009). Navigating two cultures: An investigation of cultures of a responsibility-based physical activity program and school. *Research Quarterly for Exercise and Sport*, *80*, 230–240.

Lewis, E., Ross, S., & Alberg, M. (1999). *Reliability analysis for school observation measure*. Memphis, TN: Center for Research in Educational Policy.

Li, W., Wright, P. M., Rukavina, P., & Pickering, M. (2008). Measuring students' perceptions of personal and social responsibility and its relationship to intrinsic motivation in urban physical education. *Journal of Teaching in Physical Education*, *27*, 167–178.

Martinek, T., Schilling, T., & Johnson, D. (2001). Evaluation of a sport and mentoring program designed to foster personal and social responsibility in underserved youth. *The Urban Review*, *33*, 29–45.

McKenzie, T. L., Cohen, D. A., Sehgal, A., Williamson, S., & Golinelli, D. (2006). System for observing play and recreation in communities (SOPARC): Reliability and feasibility measures. *Journal of Physical Activity and Health*, *3* (Suppl. 1), S208–S222.

McKenzie, T. L., Marshall, S. J., Sallis, J. F., & Conway, T. L. (2000). Leisure-time physical activity in school environments: An observational study using SOPLAY. *Preventive Medicine*, *30*, 70–77.

McKenzie, T. L., Sallis, J. F., & Nader, P. R. (1991). SOFIT: System for observing fitness instruction time. *Journal of Teaching in Physical Education*, *11*, 195–205.

Metzler, M. (2005). *Instructional models for physical education*. Boston, MA: Allyn & Bacon.

National Association for Sport and Physical Education (NASPE). (2004). *Moving into the future. National standards for physical education* (2nd ed, pp. 39–43). Reston, VA: Author.

Neuendorf, K. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Parker, M. (1989). Academic Learning Time-Physical Education (ALT-PE), 1982 Revision. In *Analyzing physical education and sport instruction* (2nd ed.). Champaign, IL: Human Kinetics.

Parker, M., & Hellison, D. (2001). Teaching responsibility in physical education: Standards, outcomes, and beyond. *Journal of Physical Education, Recreation, and Dance*, *72*, 25–36.

Parker, M., Kallusky, J., & Hellison, D. (1999). High impact, low risk: Ten strategies to teach responsibility. *Journal of Physical Education, Recreation, and Dance*, *70*, 26–28.

Parker, M., & Steihl, J. (2005). Personal and social responsibility. In J. Lund & D. Tannehill (Eds.), *Standards-based physical education curriculum development* (pp. 131–153). Boston: Jones and Bartlett.

Petitpas, A. J., Cornelius, A. E., Van Raalte, J. L., & Jones, T. (2005). A framework for planning youth sport programs that foster psychosocial development. *The Sport Psychologist*, *19*, 63–80.

Riffe, D., Lacy, S., & Fico, F. (1998). *Analyzing media messages: Using quantitative content analysis in research* (pp. 81–103). Mahwah, NJ: Erlbaum

Rink, J. E. (2001). Investigating the assumptions of pedagogy. *Journal of Teaching in Physical Education*, *20*, 112–128.

Ross, S. M., Smith, L. J., Alberg, M., & Lowther, D. (2004). Using classroom observation as a research and formative evaluation tool in educational reform: The School Observation Measure. In H. C. Waxman, R. G. Thorp, & R. S. Hilberg (Eds.), *Observational research in U.S. classrooms* (pp. 144–173). Cambridge UK: Cambridge University Press.

Schilling, T. A. (2001). An investigation of commitment among participants in an extended day physical activity program. *Research Quarterly for Exercise and Sport*, *72*, 355–365.

Schilling, T. A., Martinek, T., & Carson, S. (2007). Youth leaders' perceptions of commitment to a responsibility-based physical activity program. *Research Quarterly for Exercise and Sport*, *78*, 48–60.

Sterbinsky, A., & Ross, S. (2003). *School observation measure reliability study*. Memphis, TN: Center for Research in Educational Policy.

Uebersax, J. S. (1992). A review of modeling approaches for the analysis of observer agreement. *Investigative Radiology*, *27*, 738–743.

Walsh, D. (2008). Helping youth in underserved communities envision possible futures: An extension of the teaching personal and social responsibility model. *Research Quarterly for Exercise and Sport*, *79*, 209–221.

Walsh, D., Ozaeta, J., & Wright, P. M. (2010). Transference of responsibility model goals to the school environment: Exploring the impact of a coaching club program. *Physical Education and Sport Pedagogy*, *15*, 15–28.

Watson, D. L., Newton, M., & Kim, M. (2003). Recognition of values-based constructs in a summer physical activity program. *The Urban Review*, *3*, 217–232.

Wright, P. M. (2009). Research on the teaching personal and social responsibility model: Is it really in the margins? In L. Housner, M. Metzler, P. Schempp, & T. Templin (Eds.), *Historic traditions and future directions of research on teaching and teacher education in physical education* (pp. 289–296). Morgantown, WV: Fitness Information Technology.

Wright, P. M., & Burton, S. (2008). Examining the implementation and immediate outcomes of a personal-social responsibility model program for urban high school students. *Journal of Teaching in Physical Education*, *27*, 138–154.

Wright, P. M., & Li, W. (2009). Exploring the relevance of positive youth development in urban physical education. *Physical Education & Sport Pedagogy*, *14*, 241–251.

Wright, P. M., Li, W., Ding, S., & Pickering, M. (2010). Integrating a personal-social responsibility program into a lifetime wellness course for urban high school students: Assessing implementation and educational outcomes. *Sport, Education, and Society*, *15*, 277–298.