RESEARCH QUARTERLY FOR EXERCISE AND SPORT
https://doi.org/10.1080/02701367.2017.1377485

Routledge
Taylor & Francis Group

SPECIAL TOPICS: THE 2017 C. H. McCLOY LECTURE

The Intersections of Science and Practice: Examples From FitnessGram® Programming

Gregory J. Welk

Iowa State University

ABSTRACT

The FitnessGram® program has provided teachers with practical tools to enhance physical education programming. A key to the success of the program has been the systematic application of science to practice. Strong research methods have been used to develop assessments and standards for use in physical education, but consideration has also been given to ensure that programming meets the needs of teachers, students, parents, and other stakeholders. This essay summarizes some of these complex and nuanced intersections between science and practice with the FitnessGram® program. The commentaries are organized into 5 brief themes: science informing practice; practice informing science; balancing science and practice; promoting evidence-based practice; and the integration of science and practice. The article draws on personal experiences with the FitnessGram® program and is prepared based on comments shared during the 37th Annual C. H. McCloy Research Lecture at the 2017 SHAPE America – Society of Health and Physical Educators Convention.

KEYWORDS

Physical education; physical fitness; school; youth

An inherent goal of education, public health, and medical research is to promote adoption of evidence-based curricula, methods, and treatments in practice. Within medicine, researchers and clinicians receive the same basic training in medical school and share more of a common knowledge base. However, in community and public health settings, there is typically a bigger communication gap between researchers and practitioners. Recent recommendations have strongly endorsed notions of "*science with practice*" to help systematically advance public health (Galea & Annas, 2016), which has led to greater focus on external validity in research instead of solely on internal validity (Green & Glasgow, 2006; Mercer, DeVinney, Fine, Green, & Dougherty, 2007).

Scientists generally aspire to have their work and ideas adopted and used in practice. We publish in scientific journals and share our recommendations in hopes that the ideas get adopted and used, but the reality is that evidence-based programming rarely gets disseminated, effectively implemented, and sustained in real-world community settings to impact health (Remington & Brownson, 2011). As an applied kinesiology/public health researcher, I have been both intrigued and frustrated by the inherent challenges in moving the needle on population health. In this article, I will share perspectives on the intersections of science and practice that I have experienced through my own professional (and research) experiences with the FitnessGram® program. I am incredibly honored to provide these comments in recognition of the legacy of C. H. McCloy, whose pioneering work clearly advanced both research and practice in physical education and exercise science.

I will not attempt to match Brad Cardinal's (2015) excellent historical summary of C. H. McCloy's work, but I will share a few anecdotes. First, I am pleased to share a personal Iowa history with Dr. McCloy who spent the majority of his career at the University of Iowa. I spent formative years at the same institution as a young scholar pursuing my master's of science degree. Like Dr. McCloy, I have also spent the majority of my professional academic career in Iowa (but as a Cyclone at Iowa State University instead of a Hawkeye). Lastly, I also share his interests in measurement research and particularly with applied fitness assessment issues in physical education. One of McCloy's most significant contributions was his work in developing and validating a prominent motor skills assessment battery, which became known as the Iowa Brace Test (McCloy, 1937). The assessment was widely used in school settings to evaluate prerequisites to learning various motor skills taught in physical education classes

CONTACT Gregory J. Welk gwelk@iastate.edu Department of Kinesiology, Iowa State University, 103E Forker Building, 534 Wallace Avenue, Ames, IA 50011.

at the time. The list of factors mentioned in the paper included muscular strength, dynamic energy, ability to change direction, flexibility, agility, and a 'lack of an undue amount of excess fat' (McCloy & Young, 1954). The assessment also included other attributes that captured the construct of "motor educability," the 'ability to learn motor skills easily and well' (McCloy & Young, 1954). Approaches and paradigms in fitness assessment have changed considerably over time, but assessments and feedback remain the cornerstone of physical education today.

As a link to Dr. McCloy's legacy with the Iowa Brace Test, I have chosen to focus this manuscript on my own personal experiences with youth fitness assessment research and particularly with the FitnessGram® program. I was fortunate to get involved with the FitnessGram® program early in my career and have appreciated the chance to contribute as the scientific director of the program for more than 20 years. In this role, I have had opportunities to learn from many giants in the field who have launched and supported FitnessGram® over the years (see Table 1). I am particularly indebted to Dr. Chuck Corbin for his mentorship during my PhD program and throughout my career. I was steeped in underlying philosophies of fitness education (and their application within FitnessGram®) while completing my PhD, and it provided a foundation for me during the 4 years I spent working at the Cooper Institute (CI) as the director of childhood and adolescent health. I have enjoyed continuing to work with members of the FitnessGram® Scientific Advisory Board (SAB) and the staff at the CI to continue to apply science to enhance the practice of physical education programming. It has truly been one of the most satisfying aspects of my career.

There are many stories worth telling in this journey, but I have organized the article into five brief vignettes that cover different types of intersections between research and practice: (a) science informing practice, (b) practice informing science, (c) balancing science and practice, (d) promoting evidence-based practice, and (e) integration of science and practice. I generated insights about each topic through large collaborative projects in which I was involved along with various colleagues and collaborators (from the CI and on the SAB). My contributions to these projects were also directly facilitated and enhanced by many PhD students who worked with me on these lines of applied FitnessGram® research. Specific credit goes to Dr. Kelly Laurson, Dr. Pedro Saint-Maurice, and Dr. Yang Bai because they clearly taught me as much as I could have possibly taught them through these projects.

Table 1. FitnessGram® Scientific Advisory Board (past/present).

Current Board Members	Emeritus Board Members
• Joe Eisenmann	• Steve N. Blair
• Scott Going	• Charles B. Corbin
• Kathleen Janz	• Kirk J. Cureton
• Dolly Lambdin	• Harold B. Falls, Jr.
• Matt Mahar	• Timothy G. Lohman
• Jim Morrow	• Robert P. Pangrazi
• Sharon Plowman	• Russell R. Pate
• Stephen Pont	• Sarajane Quinn
• Georgi Roberts	• Margaret J. Safrit
• Gregory Welk	• James F. Sallis
• Weimo Zhu	• Charles Sterling

Although the projects and research were highly collaborative, the reflections and perspectives related to the work are my own. Therefore, the comments may not directly reflect opinions or views of others who were involved in these same projects. I chose to comment on these issues because they help to convey the complexity of applied school and public health research. I am personally committed to work at this intersection between science and practice and appreciate the opportunity to share my FitnessGram® experiences in this realm in honor of Dr. C. H. McCloy's related work and his overall legacy. In many ways, his pioneering work may have somewhat paved the way for FitnessGram® in undefined ways.

Science informing practice

The vision of a "fitness report card" by leaders with the CI provided the initial spark to set the FitnessGram® program into motion. Pioneering work by the original members of the FitnessGram® SAB brought this vision to reality (Plowman et al., 2006). A fundamental philosophical step in the development process was the decision to base the report on health-related, criterion-referenced standards. Age- and gender-specific standards were established for all the original FitnessGram® assessment items to reflect the levels of fitness needed for good health. Feedback messages were then established to enable personalized reports to be autogenerated within the FitnessGram® software. This defining feature of FitnessGram® has been instrumental in promoting effective fitness education and in advancing the practice of physical education in schools (Welk, 2006).

The original standards served the program well, but ongoing work by the members of the SAB has contributed to further refinement over the years. Setting appropriate criterion-referenced standards for fitness is an inherently complex measurement challenge—particularly in youth (Zhu, Mahar, Welk, Going, & Cureton, 2011). First of all, it is essential to have appropriate health-related indicators that are related to each of the specific fitness variables. Changes with growth

and maturation must also be taken into account to ensure that the standards work for a full K–12 physical education program. Lastly, care must be used to consider the relative implications of misclassification (i.e., false positives and false negatives). The original standards were established using sound methods, but the availability of nationally representative data from the National Health and Nutrition Examination Survey (NHANES) led to targeted efforts to refine the standards for aerobic fitness and body composition (both body fat and body mass index [BMI]), because these indicators were available on a national sample of U.S. children.

A multiyear, iterative process involving a number of experts led to refinement of standards for both aerobic fitness and body composition (Welk, Going, Morrow, & Meredith, 2011). Standards for both fitness indicators were established by examining the risk for metabolic syndrome because it is a major predictor of diabetes and cardiovascular disease. Age and maturation effects were first modeled using established LMS procedures (Eisenmann, Laurson, & Welk, 2011; Laurson, Eisenmann, & Welk, 2011a). Receiver operator characteristic (ROC) curves were then used to determine thresholds that most optimally discriminate levels of risk for metabolic syndrome (Laurson, Eisenmann, & Welk, 2011b, 2011c; Welk, Laurson, Eisenmann, & Cureton, 2011). The methods and analytic outcomes are fully described, so my interest in explaining it here is primarily to describe how the findings were operationalized and ultimately used to inform fitness education in practice.

Decisions about health standards are ultimately influenced by the relative importance of sensitivity and specificity, so considerable care was taken to establish standards that would provide appropriate feedback messages to youth about their health status. Rather than establishing a single fit–unfit standard, the decision was made to establish two different threshold values for both aerobic and body composition. In this case, the Healthy Fitness Zone (HFZ) threshold ("low risk") emphasized sensitivity because the goal was to denote sufficient fitness for health. Based on the diagnostics of the ROC method, children above this threshold have a low risk for metabolic syndrome. In contrast, the Needs Improvement Zone threshold ("high risk") emphasized specificity because it was important to avoid alarming youth and parents about low fitness unless warranted by the results. Based on the selected criteria (e.g., Sp > 95%), 95% of children without metabolic syndrome would have fitness levels above this threshold. Youth who score between the HFZ and the Needs Improvement (NI)-high risk zone are classified into an intermediate buffer zone labeled as "NI-some risk." The advantage of the three zones is that it enables more prescriptive feedback based on the documented risk potential.

The revised, empirically derived standards provided a more robust model to capture the age and gender differences in the relationship between fitness and health. Interestingly, the HFZ standards derived through this process corresponded closely with the original standards established 20 years previously without the access to nationally representative data or contemporary analytic techniques (Cureton & Mahar, 2014). However, a key programming advantage of the revised standards is that they enabled fitness levels to be categorized into three different zones instead of two. This categorization provided a way to deliver more refined fitness messages in the FitnessGram® reports. This example nicely illustrates the application of science to practice because the analytic methods were designed to specifically address this need. The next section shows how data collected through FitnessGram® has synergistically helped to advance research on youth fitness and physical education.

Practice informing science

Systematic collection and tracking of physical fitness is a common, if not defining, characteristic within physical education (Morrow, Zhu, Franks, Meredith, & Spain, 2009). The FitnessGram® has provided teachers with robust tools to collect and monitor fitness in youth, as well as guidelines for appropriate and inappropriate uses. Many large districts and states have systematically adopted FitnessGram® to standardize programming and assessment methods. Although controversial, several important commentaries previously published in *Research Quarterly for Exercise and Sport* (*RQES*) by Morrow (2005) and Morrow and Ede (2009) have emphasized that the evaluation of patterns and trends serves an important public health role in advancing the science of youth fitness. The standardized assessments provide a way to systematically monitor patterns and trends in youth fitness, to evaluate policies and programming, and to build awareness and advocacy for prevention and health promotion efforts. Through my work with the CI, I have had opportunities to contribute to ongoing evaluation of state-level data collected in both Texas and Georgia and will provide a few examples of how (in this case) practice informs science.

My initial involvement in this work came about through a large, multi-institutional collaborative research project conducted in partnership with the CI to evaluate the baseline data from the Texas Youth

Fitness Initiative (Morrow, Martin, Welk, Zhu, & Meredith, 2010). The Texas mandate required annual testing in all grades, and an ancillary grant from the Robert Woods Johnson Foundation enabled our team to systematically summarize and analyze key outcomes from this standardized data collection. The results were fully summarized in a series of manuscripts, but a few highlights are included below to illustrate the ways in which practice (i.e., school-based fitness assessment) has informed science.

- A study by Morrow et al. (2010) directly evaluated the reliability and validity of teacher-administered FitnessGram® assessments. Using a large sample and a counterbalanced design, the team directly compared the reliability of both teacher- and expert-administered tests as well as convergent validity. The reliabilities were very good to generally acceptable for all FitnessGram® tests. The validity of teacher-administered tests was also good and unrelated to school characteristics. This conclusion was important because it documents that trained teachers can conduct valid and reliable assessments.
- A study by Welk, Meredith, Ihmels, and Seeger (2010) summarized the age, gender, and regional distributions and patterns of fitness data across the state. The report was primarily descriptive, but the application of Geodemographic Information Systems tools made it possible to examine regional distributions and patterns within the state. By overlaying maps of socioeconomic status (and other variables), it was possible to visually examine social determinants of health that may explain the variability. This type of outcome can only be examined with large and systematic surveillance-type applications. Although the public health service and Centers for Disease Control and Prevention (CDC) have robust surveillance systems in place, there is still value in analyzing youth fitness data at this level to advance science (and potentially policy).
- A study by Zhu, Boiarskaia, Welk, and Meredith (2010) sought to more quantitatively understand factors that may explain variability in fitness results. Hierarchical linear modeling identified some key correlates in school physical education programs and policies that explained students' fitness status and cross-grade differences. However, the longer-term value of this study is that it documented methods for systematically evaluating school-level factors that may explain variability in fitness outcomes.

These are just examples of applied school-based evaluations that have helped to advance the science of youth fitness. While these were conducted as formalized research projects, the work would not have been possible without systematic evaluation of real-world data collected by teachers. Other ongoing studies of large-scale fitness evaluations have continued to refine methods for more effective data-processing methods. For example, in separate surveillance-related research, our team demonstrated that the assumptions used in screening and data processing of teacher-collected data have a profound effect on the outcomes and interpretation of large-scale evaluations (Saint-Maurice, Welk, Bai, & Allums-Featherston, 2014). These considerations are built into the robust sampling and analytic steps used in more controlled public health surveillance systems managed by the National Health Information System and/or CDC, but the results of this study demonstrated that care is needed to systematically evaluate large-scale fitness data collected in physical education settings.

Our team has continued to evaluate longitudinal data from Texas as well as Georgia through agreements with respective state education agencies. Annual results have been shared with key stakeholders to promote awareness for physical education programming. A surveillance study on more than 2.5 million Texas youth reported on school- and county-level correlates that have contributed to disparities in childhood obesity (Bai, Saint-Maurice, Welk, Allums-Featherston, & Candelaria, 2016). A more comprehensive longitudinal evaluation on the Georgia data provided insights into state-level patterns and trends in youth fitness (Bai & Welk, 2017). The analyses modeled 3-year (2012–2014) trajectories in HFZ achievement for both aerobic fitness and BMI after controlling for socioeconomic status and enrollment. Average annual gains in aerobic fitness achievement were 0.10% in boys and 0.87% for girls, and parallel gains in BMI achievement were 0.83% and 0.61%, respectively. Although these changes are relatively modest, they represent rather large shifts in fitness when considering the overall populations in the study. The fact that the patterns were consistent across age and gender support the premise that these are real shifts and are not due to random effects. Interestingly, we have observed similar population shifts in our related work in Texas. In Figure 1, age and gender patterns are shown for both Georgia (top) and Texas (bottom). The small but consistent trends are noteworthy from a number of perspectives, but without additional data, it is not possible to explain the causes of these favorable patterns.

The real take-home point in this section is that the data from these state adoptions have made it possible to begin to ask (and evaluate) factors that may explain

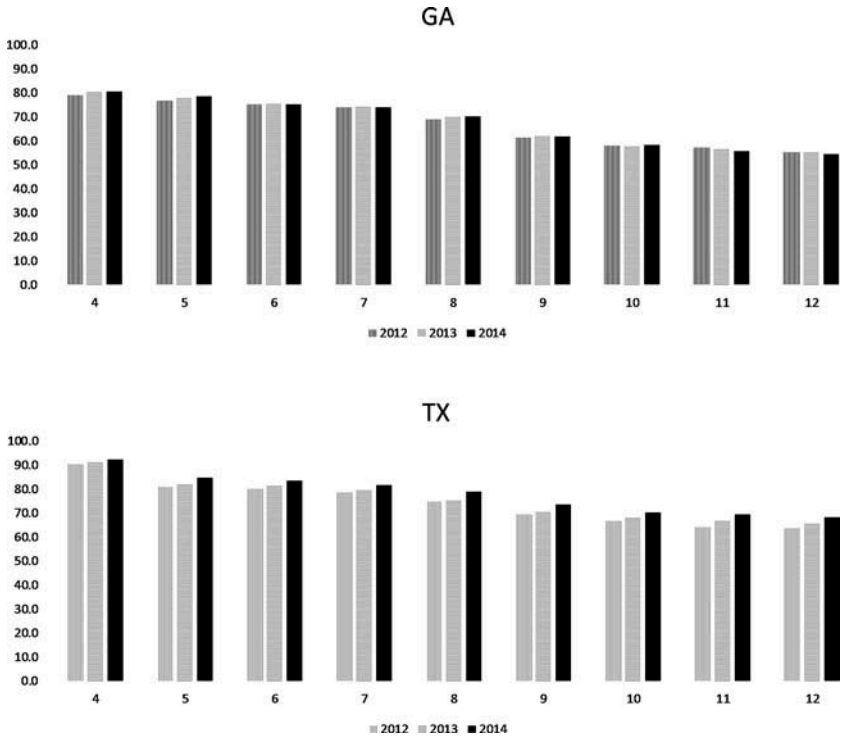

Figure 1. Statewide patterns and trends in aerobic capacity Healthy Fitness Zone achievements (boys).

differences in (and changes in) fitness and health indicators. Another important observation from these state-level data is that these patterns could not have been detected without standardized processing methods. The previously described variability in outcomes associated with choice of screening and cleaning methods (Saint-Maurice, Welk, Bai, et al., 2014) is far larger than the relatively small changes reported here. Thus, the error would have obscured the small but consistent patterns reported here if standardized methods were not used. This section summarized how evaluation of practice-based data can advance research and support ongoing advocacy efforts for physical education. The next section focuses on how research and science needs must be balanced with practice.

Balancing science and practice in youth fitness

The FitnessGram® program has sought to apply science to improve practice, but compromise has been needed to philosophically balance the relative views/needs of researchers and teachers. For example, considerable care is used by the SAB and the CI to determine whether new evidence warrants changes in the assessment batteries, the protocols, or the scoring procedures. Another interesting example of balancing research and practice came about through efforts to facilitate understanding and scoring of results from different indicators of aerobic fitness in FitnessGram®.

Schools routinely use "look-up charts" to help kids know what times they need to run in the mile (or how many Progressive Aerobic Cardiovascular Endurance Run [PACER] laps are needed) to achieve the HFZ. New charts were needed after the standards were updated because the values were expressed in maximal oxygen consumption (VO_2max) instead of mile times or laps (Welk, Laurson, et al., 2011). However, in creating the charts, it became evident that two kids could run the same time on the mile (or run the same number of laps on the PACER) and receive different estimated VO_2max levels based on their individual BMI. This finding is understandable from a research perspective because the regression equation used to predict aerobic capacity has a negative beta weight term for BMI. However, it proved hard to explain to teachers and kids (and to most everyone else). It is exceedingly difficult to reconcile why a child classified as overweight or obese would receive a lower VO_2max estimate (and a worse fitness evaluation) if they could run a mile in the same time (or achieve the same number of laps) as a child classified as normal weight. The "problem," however, is not really new. In fact, it is endemic in general classifications of aerobic fitness because VO_2 data are traditionally expressed as kilograms of body weight rather than kilograms of fat-free mass. More

detailed explanations are provided in an article by Cureton and Mahar (2014), but the emphasis here is on how the issue was ultimately addressed in the FitnessGram® to better serve the needs of teachers and students.

To resolve the problem, a new prediction equation was developed for the PACER that only includes terms for age and laps (CI, 2014). This equation allowed fitness to be evaluated based on actual performance, an approach specifically recommended by the Institute of Medicine in a prominent report (Institute of Medicine, 2012). The change in PACER scoring provided a more appropriate way to evaluate individual aerobic fitness in FitnessGram® because the scoring is independent of BMI. It has also been well received by teachers because it is simpler to score and easier to explain. Established test-equating procedures for aerobic fitness can still be used in research applications because previous research has demonstrated that mile run times and PACER laps can be interconverted with reasonable accuracy (Saint-Maurice, Welk, Laurson, & Brown, 2014; Saint-Maurice, Anderson, Bai, & Welk, 2016). However, a new mile run equation has now been developed to enable the mile run to also be scored without BMI in the FitnessGram® software (Burns et al., 2015). The key point in this vignette is that practical issues related to fitness education outweighed standard criteria commonly used to decide on appropriate prediction equations in research. The inclusion of BMI in an aerobic capacity prediction equation is clearly justifiable if the goal is to reduce the *average* error of estimation for groups of children. However, within FitnessGram®, a more important consideration is to have methods and results that are defensible at the individual level because the focus is on fitness education and personalized fitness reports. Researchers can still apply other equations with exported data, but the individual FitnessGram® reports are now processed with approaches that are more defensible (and interpretable) for children, parents, and teachers.

This section highlighted the careful balance needed between research and practice, but there are also inherent challenges in promoting evidence-based practice, which are discussed in the next section.

Promoting evidence-based practice

There is considerable variability in school physical education programming as well as with the ways in which schools and teachers use FitnessGram® data. A variety of methods and communication tools are available to promote best practices in physical education (e.g., conferences, articles, workshops, continuing education courses, Webinars, blogs etc.). FitnessGram® has worked to promote effective use and has published appropriate and inappropriate uses of fitness assessments (see Table 2), but it remains challenging to promote adoption of best practices on a large scale. The formal adoption of the FitnessGram® as the default national fitness battery by the President's Council on Physical Fitness and Sports (PCPFS) has contributed to standardization, but the distributed nature of education has led to considerable variability in utilization within districts and between states. Promising state models such as the Georgia SHAPE[1] Initiative, as described by Mike Metzler and colleagues in a 2016 McCloy Lecture (Metzler, 2016), provide examples of effective large-scale training efforts. However, more work is needed to promote consistent adoption of best practices in schools. A barrier in this regard is that methods and strategies for effective professional development are not readily shared or systematically studied across districts, states, or countries. There are no easy solutions, but a brief story is appropriate here to share a particularly impressive example of promoting evidence-based fitness education practices in Hungary.

The project evolved through a partnership between the CI and the Hungarian School Sports Federation (HSSF), which operates similarly to the PCPFS because it operates as a government-affiliated agency focused on supporting school physical education. Through a rather fortuitous set of calls and communications set

Table 2. Guidelines for appropriate and inappropriate use of fitness testing data (Position Statement from the FitnessGram® Scientific Advisory Board).

Appropriate Uses for FitnessGram®/ACTIVITYGRAM:

- Teaching students about different types of intensities of physical activity
- Teaching students about criterion-referenced health standards and health-related fitness
- Personal testing to evaluate physical activity and/or health-related fitness
- Helping students to self-monitor physical activity and track fitness results over time
- Sharing results with parents to promote family involvement and engagement
- Institutional testing to allow teachers to view group data (for curriculum development)

Inappropriate Uses for FitnessGram®/ACTIVITYGRAM:

- Evaluating individual students in physical education (e.g., grading or state standards testing)
- Using it as a sole criterion to justify students who can "test out" of physical education
- Evaluating teacher effectiveness (e.g., teacher evaluations)
- Evaluating overall physical education quality (e.g., physical education program assessment)

[1]The SHAPE initiative in Georgia was developed from the "Student Health and Physical Education (SHAPE) Act" and is not linked directly with SHAPE America – Society of Health and Physical Educators at the state level.

up by a colleague here in the United States, I connected colleagues from the HSSF with leaders at the CI. A brief in-person meeting led to an agreement to initiate an ambitious "research to practice" plan to develop a customized version of the FitnessGram® for the entire country of Hungary. The research findings proved interesting; however, the real (and heretofore, untold) story was in the *process* of coordinating the project and operationalizing the science to systematically advance the practice of physical education in the country.

With modest consulting support, the HSSF team first coordinated the collection of field and lab data needed to evaluate the utility of the FitnessGram® standards in a large and representative sample of Hungarian youth (Csányi et al., 2015). The field-based protocols were based on standard FitnessGram® test administration manuals, and data were systematically collected on a random sample of youth across the entire country. A subsample of youth from multiple schools in each region were then recruited to participate in the lab-based data collection conducted through regional medical centers. The lab protocols were based on protocols used in NHANES because this data system was the basis for the FitnessGram® standards. Results of this work were summarized in a series of collaborative articles published by members of the CI-based research team and HSSF researchers. An independent quality-control evaluation demonstrated good fidelity to the lab and field protocols (Csányi et al., 2015), and the analyses provided support for the utility of the FitnessGram® standards in this national sample of Hungarian youth (Laurson, Saint-Maurice, Karsai, & Csányi, 2015; Laurson, Welk, Marton, Kaj, & Csányi, 2015; Saint-Maurice, Welk, Finn, & Kaj, 2015). The results provided baseline information on the distribution of health-related fitness in Hungarian youth (Welk, Saint-Maurice, & Csányi, 2015), and overall results were shared at an international conference less than 2 years later.

The coordination of this research phase was impressive enough, but the approaches used to train teachers and to institutionalize effective use of the new assessment battery provide a great example of promoting evidence-based programming in schools. The CI team worked with HSSF leaders to develop a customized Web-based platform called "*NetFit*" that was based on FitnessGram® approaches and concepts. Sample coding and feedback messages were provided to enable individual and group reports to be generated (similar to the FitnessGram®). Manuals and video-based training modules were then developed by the HSSF to facilitate training and implementation across the entire country. Extensive professional development programming (including in-person training) was ultimately carried out in a systematic manner to train teachers across the country on appropriate fitness education methods based on NetFit. There are obviously many gaps in this story, but the HSSF team essentially led a countrywide transformation from a sport-based model to a health-related fitness education approach in 2 to 3 years.

The articles documenting the utility of the standards provide an important research basis for the programming, but the operationalization of the findings into professional development training by the HSSF team in Hungary provided a good example of how to systematically implement evidence-based practices. Large-scale training and dissemination efforts are obviously more difficult in larger and more decentralized states and countries, and there are clearly many parallel efforts in the United States and other countries to accomplish the same goals. Mike Metzler described steps needed to systematically advance physical education programming in his recent McCloy Lecture on "grand challenges" (Metzler, 2016). The next section describes the advantages of integrating science and practice to help understand what works and what does not work.

Integration of science and practice

A challenge in public health research is that evidence-based programming rarely gets incorporated into standard practice (Remington & Brownson, 2011). Advancing "*science with practice*" and system-based strategies has been widely endorsed to help address this gap (Galea & Annas, 2016).

To advance understanding of the factors that influence successful physical education programming, the CI, in partnership with the NFL PLAY 60 Foundation, established the NFL PLAY 60 FitnessGram® Partnership Project. The goal of the project is to help schools take full advantage of the coordinated fitness and physical activity resources available through FitnessGram® and the NFL PLAY 60 programs. However, a novel aspect is that it was set up using participatory approaches that give schools complete autonomy over the degree to which they follow or use the recommendations. The programming first focused on building the capacity to use the FitnessGram® assessment. Once schools were able to complete and submit results, they were encouraged to work through the process of planning and implementing programming available through NFL Play 60. Schools were guided through the process in sequential cohorts to

allow them to progress at different phases (Welk et al., 2016).

Recruitment was done directly in partnership with the National Football League (NFL) clubs with each of the NFL franchises being positioned to distribute 35 site licenses to schools in their respective areas. The recruitment was slow but systematic; however, an advantage was that it generated interest and involvement from the franchises over time. The project provides key advantages to a number of stakeholders. Schools learn how to optimize the use of FitnessGram® and activity programming, the CI and NFL PLAY 60 program learn about strategies to enhance program effectiveness, and researchers learn how to most effectively support and assist schools (Welk et al., 2016).

Although not developed as a specific surveillance project, the large and distributed sample provided a useful benchmark for a national profile of health-related fitness (Bai et al., 2015) as well as evaluations of fitness disparities (Bai et al., 2016). The publication of studies from this large, participatory research network in prominent medical journals is noteworthy because it documents the acceptance of this type of field-based data collection for research applications. Earlier studies documenting the viability of teacher-administered fitness data in Texas (Morrow, Martin, & Jackson, 2010) and the development of standardized screening methods (Saint-Maurice, Welk, Bai, et al., 2014) were likely important in documenting the utility of the school-based data collection. The results filled an important gap in the literature because they documented levels of health-related fitness in a large sample of U.S. schools—findings that have not been possible due to the lack of a school-based fitness surveillance system.

The most prominent study from this ongoing project revealed longitudinal changes in fitness resulting from systematic adoption of NFL PLAY 60 programming in schools (Bai, Saint-Maurice, Welk, Allums-Featherston, & Candelaria, 2017). Use of NFL PLAY 60 was encouraged in the project but was not required. Schools were considered to be "*programming schools*" if they reported on an annual survey that they used either "*Fuel Up to PLAY 60*" or "*PLAY 60 Challenge*" at least 2 years of the 4 years of tracking. With this simple classification, we were able to systematically compare longitudinal trends in fitness for schools that used NFL PLAY 60 programming to those in schools that did not. Robust growth curve models demonstrated that programming schools had larger gains in aerobic fitness HFZ achievement than did nonprogramming schools. The analyses controlled for school and teacher characteristics such as school health policy, physical education duration, and teacher teaching experience as well as socioeconomic status. The mean annual gains (3% in girls and 2.9% for boys) reflect substantial improvements, particularly when one considers that schools were provided with only recommendations to use programming and did not receive any specific support or funding to carry it out. Gains in BMI HFZ were smaller (~1%) but still noteworthy considering that many school-based interventions do not observe any changes in BMI.

The key strength of this project was the participatory nature of the design because it enabled the findings to reflect outcomes that are more "real-world." Annual surveys and teacher feedback provide insights about the needs and motivations of teachers in a diverse sample of schools. Instead of evaluating outcomes that might be possible under perfect conditions, the results reflect more generalizable outcomes that are possible with implementation under real-world conditions. Ultimately, the goal of the NFL project was to understand how to impact and improve physical education programming so that the methods can be applied to other schools to enhance physical education programming on a larger scale.

Conclusion

The FitnessGram® program has had a long and lasting impact on the evolution and growth of physical education in schools (Plowman et al., 2006). The strong, positive legacy can be attributed, in large part, to a continued emphasis on (and adherence to) science. However, close attention to the specific needs of practitioners and other key stakeholders has also been essential. The commentaries about my experiences with the FitnessGram® program were intended to capture some of the intersections and interactions that operate between science and practice. My depictions only hint at the complexities, but Larry Green's famous mantra provides a good summation: "*If you want evidence-based practice, you need to have to have practice-based evidence*" (Green, 2006, p. 406).

As an applied kinesiology/public health researcher, I have enjoyed the opportunity to contribute to the growth and development of the FitnessGram® program for the past 20 years. I have been fortunate to have had opportunities to work with many leaders in the field, and I am honored to provide these reflections and comments on behalf of my colleagues and students who have worked with me on this front. Dr. C. H. McCloy's legacy provides continued inspiration to guide our collective efforts to carefully and intentionally integrate science with practice.

What does this article add?

The commentary provides perspectives about the complex interplay between science and practice. The examples provided here were specific to FitnessGram®, but similar issues are at play in other areas, industries, and fields. The comments in the manuscript were based on insights and observations from a number of large, multi-year collaborative projects. While the published articles resulting from these projects provide important conclusions, other valuable insights and lessons often do not get reported or discussed. Thus, a goal in the manuscript was to share some of these hidden perspectives.

ORCID

Gregory J. Welk http://orcid.org/0000-0001-7132-9725

References

Bai, Y., & Welk, G. J. (in press). Longitudinal youth fitness trends and disparity in the state of Georgia. *Public Health Reports*.

Bai, Y., Saint-Maurice, P. F., Welk, G. J., Allums-Featherston, K., Candelaria, N., & Anderson, K. (2015). Prevalence of youth fitness in the United States: Baseline results from the NFL PLAY 60 FitnessGram® Partnership Project. *Journal of Pediatrics*, *167*, 662–668. doi:10.1016/j.jpeds.2015.05.035

Bai, Y., Saint-Maurice, P. F., Welk, G. J., Allums-Featherston, K. A., & Candelaria, N. (2016). Explaining disparities in youth aerobic fitness and body mass index: Relative impact of socioeconomic and minority status. *Journal of School Health*, *86*, 787–793. doi:10.1111/josh.2016.86.issue-11

Bai, Y., Saint-Maurice, P. F., Welk, G. J., Allums-Featherston, K. A., & Candelaria, N. (2017). The longitudinal impact of NFL PLAY 60 programming on youth aerobic capacity and body mass index. *American Journal of Preventive Medicine*, *52*, 311–323. doi:10.1016/j.amepre.2016.10.009

Bai, Y., & Welk, G. (2017). Multilevel analysis of school and county correlates associated with youth body mass index. *Medicine & Science in Sports & Exercise*, *49*, 1842–1850. doi:10.1249/MSS.0000000000001311

Burns, R. D., Hannon, J. C., Brusseau, T. A., Saint-Maurice, P. F., Welk, G. J., & Mahar, M. T. (2015). Cross-validation of aerobic capacity prediction models in adolescents. *Pediatric Exercise Science*, *27*, 404–411. doi:10.1123/pes.2014-0175

Cardinal, B. J. (2015). The 2015 CH McCloy Lecture: Road trip toward more inclusive physical activity: Maps, mechanics, detours, and traveling companions. *Research Quarterly for Exercise and Sport*, *86*, 319–328.

Cooper Institute. (2014). *FitnessGram®/ACTIVITYGRAM reference guide* (4th ed., S. A. Plowman & M. D. Meredith, Eds.) Dallas, TX: Author.

Csányi, T., Finn, K. J., Welk, G. J., Zhu, W., Karsai, I., Ihász, F., … Molnár, L. (2015). Overview of the Hungarian National Youth Fitness Study. *Research Quarterly for Exercise and Sport*, *86*(Suppl. 1), S3–S12. doi:10.1080/02701367.2015.1042823

Cureton, K. J., & Mahar, M. T. (2014). Critical measurement issues/challenges in assessing aerobic capacity in youth. *Research Quarterly for Exercise and Sport*, *85*, 136–143. doi:10.1080/02701367.2014.898979

Eisenmann, J. C., Laurson, K. R., & Welk, G. J. (2011). Aerobic fitness percentiles for U.S. adolescents. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S106–S110.

Galea, S., & Annas, G. J. (2016). Aspirations and strategies for public health. *Journal of the American Medical Association*, *315*, 655–656. doi:10.1001/jama.2016.0198

Green, L. W. (2006). Public health asks of systems science: To advance our evidence-based practice, can you help us get more practice-based evidence? *American Journal of Public Health*, *96*, 406–409. doi:10.2105/AJPH.2005.066035

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. *Evaluation & the Health Professions*, *29*, 126–153. doi:10.1177/0163278705284445

Institute of Medicine of the National Academies. (2012). *Fitness measures and health outcomes in youth*. Washington, DC: National Academies Press.

Laurson, K. R., Eisenmann, J. C., & Welk, G. J. (2011a). Body fat percentile curves for U.S. children and adolescents. *American Journal of Preventive Medicine.*, *41*(Suppl. 2), S87–S92. doi:10.1016/j.amepre.2011.06.044

Laurson, K. R., Eisenmann, J. C., & Welk, G. J. (2011b). Development of BMI standards based on measurement agreement with health related body fat standards. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S100–S105. doi:10.1016/j.amepre.2011.07.004

Laurson, K. R., Eisenmann, J. C., & Welk, G. J. (2011c). Development of youth percent body fat standards using receiver operating characteristic curves. *American Journal of Preventive Medicine.*, *41*(Suppl. 2), S93–S99. doi:10.1016/j.amepre.2011.07.003

Laurson, K. R., Saint-Maurice, P. F., Karsai, I., & Csányi, T. (2015). Cross-validation of FitnessGram® health-related fitness standards in Hungarian youth. *Research Quarterly for Exercise and Sport*, *86*(Suppl. 1), S13–S20. doi:10.1080/02701367.2015.1042800

Laurson, K. R., Welk, G. J., Marton, O., Kaj, M., & Csányi, T. (2015). Agreement and diagnostic performance of FitnessGram®, International Obesity Task Force, and Hungarian national BMI standards. *Research Quarterly for Exercise and Sport*, *86*(Suppl. 1), S21–S28. doi:10.1080/02701367.2015.1042786

McCloy, C. H. (1937). An analytical study of the stunt type test as a measure of motor educability. *Research Quarterly*, *8*, 46–55.

McCloy, C. H., & Young, N. D. (1954). *Tests and measurements in health and physical education*. New York, NY: Appleton-Century-Crofts.

Mercer, S. L., DeVinney, B. J., Fine, L. J., Green, L. W., & Dougherty, D. (2007). Study designs for effectiveness and translation research: Identifying trade-offs. *American Journal of Preventive Medicine*, *33*, 139–154. doi:10.1016/j.amepre.2007.04.005

Metzler, M. W. (2016). School-based team research to address grand challenges through P–12 physical education programs. *Research Quarterly for Exercise and Sport*, *87*, 325–333. doi:10.1080/02701367.2016.1234284

Morrow, J. R., Jr. (2005). 2004 CH McCloy Research Lecture: Are American children and youth fit? It's time we learned. *Research Quarterly for Exercise and Sport*, *76*, 377–388.

Morrow, J. R., Jr., & Ede, A. (2009). *Research Quarterly for Exercise and Sport* Lecture. Statewide physical fitness testing: A big waist or a big waste? *Research Quarterly for Exercise and Sport*, *80*, 696–701.

Morrow, J. R., Jr., Martin, S. B., & Jackson, A. W. (2010). Reliability and validity of the FitnessGram®: Quality of teacher-collected health-related fitness surveillance data. *Research Quarterly for Exercise and Sport*, *81*(Suppl. 3), S24–S30. doi:10.1080/02701367.2010.10599691

Morrow, J. R., Jr., Martin, S. B., Welk, G. J., Zhu, W., & Meredith, M. D. (2010). Overview of the Texas Youth Fitness Study. *Research Quarterly for Exercise and Sport*, *81*(Suppl. 3), S1–S5. doi:10.1080/02701367.2010.10599688

Morrow, J. R., Jr., Zhu, W., Franks, D. B., Meredith, M. D., & Spain, C. (2009). 1958–2008: 50 years of youth fitness tests in the United States. *Research Quarterly for Exercise and Sport*, *80*, 1–11.

Plowman, S. A., Sterling, C. L., Corbin, C. B., Meredith, M. D., Welk, G. J., & Morrow, J. J. R. (2006). The history of FitnessGram®. *Journal of Physical Activity and Health*, *3* (Suppl. 2), S5–S20. doi:10.1123/jpah.3.s2.s5

Remington, P. L., & Brownson, R. C. (2011). Fifty years of progress in chronic disease epidemiology and control. *Morbidity and Mortality Weekly Report*, *60*(4), 70–77.

Saint-Maurice, P. F., Anderson, K., Bai, Y., & Welk, G. J. (2016). Agreement between VO2 peak predicted from PACER and one-mile run time-equated laps. *Research Quarterly for Exercise and Sport*, *87*, 421–426. doi:10.1080/02701367.2016.1216067

Saint-Maurice, P. F., Welk, G. J., Bai, Y., & Allums-Featherston, K. (2014). Comparison of screening methods for evaluating school-level fitness patterns with FitnessGram®: Findings from the NFL PLAY 60 FitnessGram® partnership. *Open Journal of Preventive Medicine*, *4*, 876–886. doi:10.4236/ojpm.2014.411099

Saint-Maurice, P. F., Welk, G. J., Finn, K. J., & Kaj, M. (2015). Cross-validation of a PACER prediction equation for assessing aerobic capacity in Hungarian youth. *Research Quarterly for Exercise and Sport*, *86*(Suppl. 1), S66–S73. doi:10.1080/02701367.2015.1043002

Saint-Maurice, P. F., Welk, G. J., Laurson, K. R., & Brown, D. D. (2014). Measurement agreement between estimates of aerobic fitness in youth: The impact of body mass index. *Research Quarterly for Exercise and Sport*, *85*, 59–67. doi:10.1080/02701367.2013.872217

Welk, G. J. (2006). Strengthening the scientific basis of the *FitnessGram*® program. *Journal of Physical Activity and Health*, *3*(Suppl. 2), S1–S4. doi:10.1123/jpah.3.s2.s1

Welk, G. J., Bai, Y., Saint-Maurice, P. F., Candeleria, N., Allums-Featherston, K. A., & Anderson, K. (2016). Design and evaluation of the NFL PLAY 60 FitnessGram® Partnership Project. *Research Quarterly for Exercise and Sport*, *87*, 1–13. doi:10.1080/02701367.2015.1127126

Welk, G. J., Laurson, K. R., Eisenmann, J. C., & Cureton, K. J. (2011). Development of youth aerobic capacity standards using receiver operator characteristic curves. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S111–S116. doi:10.1016/j.amepre.2011.07.007

Welk, G. J., Going, S. B., Morrow, J. R., Jr., & Meredith, M. D. (2011). Development of new criterion-referenced fitness standards in the FitnessGram® program: Rationale and conceptual overview. *American Journal of Preventive Medicine*, *41*(Suppl. 2), S63–S68. doi:10.1016/j.amepre.2011.07.012

Welk, G. J., Meredith, M. D., Ihmels, M., & Seeger, C. (2010). Distribution of health-related physical fitness in Texas youth: A demographic and geographic analysis. *Research Quarterly for Exercise and Sport*, *81*(Suppl. 3), S6–S15. doi:10.1080/02701367.2010.10599689

Welk, G. J., Saint-Maurice, P. F., & Csányi, T. (2015). Health-related physical fitness in Hungarian youth: Age, sex, and regional profiles. *Research Quarterly for Exercise and Sport*, *86*(Suppl. 1), S45–S57. doi:10.1080/02701367.2015.1043231

Zhu, W., Boiarskaia, E. A., Welk, G. J., & Meredith, M. D. (2010). Physical education and school contextual factors relating to students' achievement and cross-grade differences in aerobic fitness and obesity. *Research Quarterly for Exercise and Sport*, *8*(Suppl. 3), S53–S64. doi:10.1080/02701367.2010.10599694

Zhu, W., Mahar, M. T., Welk, G. J., Going, S. B., & Cureton, K. J. (2011). Approaches for development of criterion-referenced standards in health-related youth fitness tests. *American Journal of Preventive Medicine.*, *41*(Suppl. 4), S68–S76. doi:10.1016/j.amepre.2011.07.001

Zhu, W., Welk, G. J., Meredith, M. D., & Boiarskaia, E. A. (2010). A survey of physical education programs and policies in Texas schools. *Research Quarterly for Exercise and Sport*, *81*(Suppl. 3), S42–S52. doi:10.1080/02701367.2010.10599693