Reflecting on Field Performance Tests of Pediatric Aerobic Fitness: After 30 Years, It Really Is Time to Move on

Jo Welsman
University of Exeter

The May 2019 Special Issue is of great interest to those of us who have been reading Pediatric Exercise Science since the earliest days of the journal. We recall with great fondness the thoughtful, critical, and often provocative notes by the journal’s founder and long-term editor Professor Tom Rowland. In the very first volume, Tom alluded to the strength of emotion aroused by debates of the pros and cons of mass physical fitness testing of children and youth, encouraged active and open debate (16), and subsequently advised us that “the horse is dead; let’s dismount” (18). Exeter researchers have shared these concerns since the early 1980s and in 1988 published an evaluation of the validity of the 20-m shuttle run test (20mSRT) in 77 boys (aged 11–14 y). A common variance of 29% between 20mSRT performance and rigorously determined peak VO2 was reported. On the basis of these findings, it was concluded that the use of the 20mSRT could not be supported as a valid substitute for a direct determination of peak VO2 (4). As research techniques progressed over the next 30 years and infinitely more nuanced information on children’s aerobic (or cardiopulmonary) fitness was gleaned from longitudinal studies, breath by breath studies of oxygen uptake kinetics, and magnetic resonance imaging and spectroscopic techniques supported by allometric and multi-level regression modeling, we assumed (naively it transpires) that field performance tests would gradually recede from use in rigorous scientific research.

It is, therefore, of very deep concern that despite early recognition of the serious limitations of predictive field tests, over recent years, the uncritical use of the 20mSRT has, like an unchecked wildfire, engulfed a more scientifically rigorous approach to understanding relationships between fitness and health in youth—the flames fanned no doubt by the apparently simple virtues of the test, which are frequently and repetitiously extolled. As a result, 20mSRT data from children as young as 2 years have been converted into “reference standards for preschool children” (5); 20mSRT scores on over a million children aged from 9 to 17 years have been used to estimate peak VO2 and collate values ratio-scaled with body mass into international aerobic fitness “norms” (21); and inter-country comparisons of “who are the fittest?” based on peak VO2 predicted from 20mSRT performance have been reported (9). 20mSRT performance has been recommended to evaluate physical activity interventions (7), survey and monitor health (8), elucidate metabolic and cardiovascular risk (6), and identify individual children who warrant specific intervention to improve current and future health and fitness—the so-called clinical red flags (19).

However, just because a test is widely used, inexpensive, and (perhaps the key driver) able to accumulate data on numerous participants quickly and with minimal effort does not qualify it as a criterion measure for studies of youth fitness and health. None of these qualities relate to what is absolutely and unquestionably fundamental for any measurement: validity and reliability—issues that are too often, too easily skimmed over by those deploying the tests within their otherwise, and this is what is really baffling, carefully designed, and meticulously analyzed research.

The 20mSRT is undoubtedly the most popular field test of aerobic fitness, but for a test with such global reach there are remarkably few studies of criterion validity. Data are based on relatively small samples of predominantly teenage volunteers with conclusions largely based on test–retest correlation. When summarized through meta-analysis into the best evidence to date, the criterion validity of the 20mSRT in children is revealed to be only “moderate.” Mayorga-Vega et al (14) reported that over 50% of criterion validity studies with children report a shared variance of 49% or less between laboratory-determined peak VO2 and that estimated from 20mSRT.

Results explored with more informative validity statistics are rare, and based on sample sizes that are too small to be generalizable. Although some studies provide standard error of the estimate alongside their predictions, confidence intervals around these are not reported. Matsuzuka et al (13) reported 95% limits of agreement of −9.8 to 6.4 mL·kg⁻¹·min⁻¹, representing −32% of measured peak VO2. Limits of agreement have also been reported for 20mSRT-estimated peak VO2 versus the highest VO2 measured, using portable gas analysis equipment, during the test. Even so, similarly, wide 95% limits of agreement of −6 to −19 mL·kg⁻¹·min⁻¹ were noted in addition to significantly different mean values when predictions were based on the widely used Léger equation (20).

If that were not sufficient evidence to dissuade researchers from pursuing this methodology, the reliability data are more worrying. Reliability of the 20mSRT based on intraclass correlations and mean differences is reported to be “strong to very strong” (21). Much more revealing, however, are the results of test–retest statistics, based on limits of agreement: should we really consider 95% limits of agreement of ±/− 2.5 stages on a test lasting 4 to 6 stages reliable? (15). Those data were obtained in teenage volunteers. Is the test likely to be more reliable in preschool children? If 20mSRT performance is a valid estimate of peak VO2, should not we expect the reliability to be somewhere closer to that of its criterion reference?

The author is with Children’s Health and Exercise Research Centre, University of Exeter, Exeter, United Kingdom. Welsman (j.r.welsman2@exeter.ac.uk) is corresponding author.
In addition to issues with validity and reliability, and perhaps of most significance, is that the entire behemoth of 20mSRT literature is founded on an incorrect mathematical assumption—that is, that expressing peak \( \dot{VO}_2 \) in simple ratio with body mass is (1) an accurate mathematical description of data and (2) sufficient to account for the changes in body size and composition that necessarily accompany growth and adolescence. Quite simply, it is not. In yet another echo from the earliest days of Pediatric Exercise Science, Tom Rowland (17) referred to the issue of normalizing for body size but to summarize briefly, in none of our extensive cross-sectional or longitudinal data sets incorporating over 2000 treadmill peak \( \dot{VO}_2 \) determinations in approximately equal numbers of boys and girls, have we ever found the simple per body mass ratio to accurately and completely normalize data. Traditional and convenient it may be, but increasingly data are demonstrating that age and maturational changes in body mass and composition are necessary concurrent covariates when interpreting youth aerobic fitness.

If we consider these 3 issues around validity, reliability, and the fact that these essential foundations are themselves founded on inappropriately expressed data, then some of the 20mSRT-based trends that conflict with our understanding of results derived from laboratory-determined aerobic fitness are perhaps explained. For example, why are prepubertal girls reported as substantially less fit than prepubertal boys when both laboratory-derived data and anthropometric and physiological characteristics suggest to us they should be within ~10%? Perhaps girls are less motivated to run 20-m shuttles? Perhaps data reflect cultural differences in girls’ willingness to exercise publically to maximal effort? (eg, 12). Why does estimated peak \( \dot{VO}_2 \) from 20mSRT decline in girls during adolescence when extensive longitudinal data tell us that there is no age-related decline in fitness once appropriately scaled for mass and fatness (2)? Perhaps they are just getting heavier and increasing body fat during adolescence—which is not the same as getting less fit! Why are overweight and obese youngsters so often classified as unfit using 20mSRT scores when rigorously determined and appropriately analyzed data suggest minimal aerobic fitness differences between normal and overweight children (2,11)? Perhaps this is a classic illustration of how ratio scaling underestimates fitness in heavier individuals? Why are temporal trends in aerobic fitness estimated from 20mSRT performance showing a decline when laboratory data are not (3)? Perhaps that is because the 20mSRT is not a valid, reliable estimate of peak \( \dot{VO}_2 \). Why are single-value clinical “cutoffs” being recommended for individual intervention for youth from 8 to 18 years of age (10,19) when large sets of laboratory-determined longitudinal data demonstrate that peak \( \dot{VO}_2 \) changes with age and maturity status (1,2)?

In summary, it is hard to dispute that the 20mSRT is a widely used, practical, and cheap test of running performance founded on a simple, convenient, and traditional method of normalizing data for body size differences. However, the 20mSRT is not an accurate estimate of peak \( \dot{VO}_2 \) (estimates within ~25%–30% of measured peak \( \dot{VO}_2 \)), reliable (values may vary by up to ~50%), or able to account for changes in body mass, muscularity, and fatness during growth. And yet it is this test that is repeatedly and extensively recommended for global use to determine, compare, and monitor levels of child and adolescent fitness; examine relationships with health indicators and disease risk; and identify individual children at “clinical risk.” Whatever happened to scientific rigor and critical analysis? In attempting to simplify relationships that are “extraordinarily complex and defy simplification” (17), we are in danger of misdirecting a generation of research and researchers, imposing inappropriate interventions on children and adolescents and failing to understand true and potentially important relationships between aerobic fitness and present and future metabolic risk. Our children deserve better than this.

References