January 2016

Towards best practice in physical and physiological employment standards

Stewart R. Petersen
University of Alberta

Gregory S. Anderson
Justice Institute of British Columbia

Michael J. Tipton
University of Portsmouth, michael.tipton@port.ac.uk

David Docherty
University of Victoria

Terry E. Graham
University of Guelph

See next page for additional authors

Towards best practice in physical and physiological employment standards

Abstract

While the scope of the term physical employment standards is wide, the principal focus of this paper is on standards related to physiological evaluation of readiness for work. Common applications of such employment standards for work are in public safety and emergency response occupations (e.g., police, firefighting, military), and there is an ever-present need to maximize the scientific quality of this research. Historically, most of these occupations are male-dominated, which leads to potential sex bias during physical demands analysis and determining performance thresholds. It is often assumed that older workers advance to positions with lower physical demand. However, this is not always true, which raises concerns about the long-term maintenance of physiological readiness. Traditionally, little attention has been paid to the inevitable margin of uncertainty that exists around cut-scores. Establishing confidence intervals around the cut-score can reduce for this uncertainty. It may also be necessary to consider the effects of practise and biological variability on test scores. Most tests of readiness for work are conducted under near perfect conditions, while many emergency responses take place under far more demanding and unpredictable conditions. The potential impact of protective clothing, respiratory protection, load carriage, environmental conditions, nutrition, fatigue, sensory deprivation, and stress should also be considered when evaluating readiness for work. In this paper, we seek to establish uniformity in terminology in this field, identify key areas of concern, provide recommendations to improve both scientific and professional practice, and identify priorities for future research.

Publication Details

Petersen, S. R., Anderson, G. S., Tipton, M. J., Docherty, D., Graham, T. E., Sharkey, B. J. & Taylor, N. A.S.. Towards best practice in physical and physiological employment standards. Applied Physiology, Nutrition, and Metabolism. 2016; 41 (6, Suppl. 2): S47-S62.

Authors

Stewart R. Petersen, Gregory S. Anderson, Michael J. Tipton, David Docherty, Terry E. Graham, Brian J. Sharkey, and Nigel A.S. Taylor

NRC Research Press

REVIEW

Towards best practice in physical and physiological employment standards[1]

Stewart R. Petersen, Gregory S. Anderson, Michael J. Tipton, David Docherty, Terry E. Graham, Brian J. Sharkey, and Nigel A.S. Taylor

Abstract: While the scope of the term *physical employment standards* is wide, the principal focus of this paper is on standards related to *physiological* evaluation of readiness for work. Common applications of such employment standards for work are in public safety and emergency response occupations (e.g., police, firefighting, military), and there is an ever-present need to maximize the scientific quality of this research. Historically, most of these occupations are male-dominated, which leads to potential sex bias during physical demands analysis and determining performance thresholds. It is often assumed that older workers advance to positions with lower physical demand. However, this is not always true, which raises concerns about the long-term maintenance of physiological readiness. Traditionally, little attention has been paid to the inevitable margin of uncertainty that exists around cut-scores. Establishing confidence intervals around the cut-score can reduce for this uncertainty. It may also be necessary to consider the effects of practise and biological variability on test scores. Most tests of readiness for work are conducted under near perfect conditions, while many emergency responses take place under far more demanding and unpredictable conditions. The potential impact of protective clothing, respiratory protection, load carriage, environmental conditions, nutrition, fatigue, sensory deprivation, and stress should also be considered when evaluating readiness for work. In this paper, we seek to establish uniformity in terminology in this field, identify key areas of concern, provide recommendations to improve both scientific and professional practice, and identify priorities for future research.

Key words: employment standards, human rights, standard setting, cut-scores, reliability, validity.

Résumé : La portée de l'expression *normes physiques relatives à l'emploi* étant vaste, cet article met l'accent sur les normes associées à l'évaluation *physiologique* de l'aptitude au travail. On utilise généralement ces normes d'emploi dans le domaine de la sécurité publique et des interventions d'urgence (p. ex. policier, pompier, militaire); ceci étant, il est toujours possible d'améliorer la qualité scientifique de ce domaine de recherche. Historiquement, la majorité de ces emplois sont occupés par des hommes, ce qui pourrait inclure un biais lié au sexe lors de l'analyse des exigences physiques et de la fixation des seuils de performance. Il semble acquis que les travailleurs plus âgés occupent au fil du temps des postes moins exigeants physiquement. Toutefois, comme ce n'est pas toujours le cas, le maintien à long terme des aptitudes au travail pourrait s'avérer problématique. Depuis longtemps, on porte peu d'attention à la marge d'incertitude concernant les seuils de coupure. L'établissement d'un intervalle de confiance au seuil de coupure pourrait combler cette incertitude. On devrait possiblement prendre en compte les effets de la pratique et de la variabilité biologique sur les résultats des tests. La plupart des évaluations des aptitudes au travail sont réalisées dans des conditions presque parfaites; pourtant, nombre d'interventions d'urgence sont effectuées dans des conditions beaucoup plus exigeantes et imprévisibles. Quand il est question d'évaluer les aptitudes à l'emploi, on devrait prendre en compte les effets potentiels des vêtements de protection, de la protection des voies respiratoires, du port de charge, des conditions environnementales, de l'alimentation, de la fatigue, de la privation sensorielle et du stress. Dans cet article, nous souhaitons installer une uniformité terminologique, identifier les principaux secteurs de préoccupation, faire des recommandations pour améliorer la pratique scientifique et professionnelle et définir les priorités pour les études ultérieures. [Traduit par la Rédaction]

Mots-clés : normes d'emploi, droits de la personne, installation de base, seuils de coupure, fiabilité, validité.

Received 4 January 2016. Accepted 31 March 2016.

S.R. Petersen. Faculty of Physical Education and Recreation, University of Alberta, Edmonton, AB T6G 2H9, Canada.
G.S. Anderson. Justice Institute of British Columbia, New Westminster, BC V3L 5T4, Canada.
M.J. Tipton. Department of Sport and Exercise Sciences, University of Portsmouth, Portsmouth, Hants, PO1 2ER, UK.
D. Docherty. School of Exercise Science, Physical & Health Education, University of Victoria, Victoria, BC V8P 5C2, Canada.
T.E. Graham.* Department of Human Health and Nutritional Sciences, University of Guelph, Guelph, ON N1G 2W1, Canada.
B.J. Sharkey. Department of Health and Human Performance, University of Montana, Missoula, MT 59812, USA.
N.A.S. Taylor. Centre for Human and Applied Physiology, School of Medicine, University of Wollongong, Wollongong, NSW 2522, Australia.

Corresponding author: Stewart R. Petersen (email: stewart.petersen@ualberta.ca).

*Terry E. Graham currently serves as an Editor; peer review and editorial decisions regarding this manuscript were handled by Glen P. Kenny.

[1]This paper is part of a supplemental issue entitled Proceedings from the Second International Conference on Physical Employment Standards – Best Practice in Physical Employment Standards: An International Perspective. Second International Conference on Physical Employment Standards (PES 2015) was held in Canmore, Alberta, Canada; August 23–26, 2015.

Introduction

The field of employment standards for physically demanding jobs is a complex intersection between disciplinary interests such as human rights, law, medicine, occupational health and safety, and physiology. The focus of these interactions is not stress avoidance, because that is not possible in many jobs. Rather, it is about defining and understanding work-related stress so that it can be minimized where possible, and workers can be recruited and trained to tolerate the unavoidable physical demands. In August 2015, scientific and professional delegates from around the world attended the *Second International Conference on Physical Employment Standards* (Canmore, Alberta, Canada) to address critical questions in the domain of employment standards. That meeting followed the first dedicated conference in 2012 (Canberra, Australia; Taylor and Billing 2012), and will be followed by a third meeting in the United Kingdom (July 2018). The program for the second conference was informed by a series of invited review papers (also published in this issue), interactive knowledge translation sessions, and original research presentations. The outcomes of that conference are the focus of this special issue, and represent, on a global scale, the current state of knowledge to guide best practice in this important field. Immediately following the conference, a smaller group of ~50 stakeholders representing the various constituencies in the field (e.g., scientists, practitioners, managers) participated in 2 days of facilitated discussion on key issues in the field that chiefly arose from those invited reviews.

The authors of the present article were tasked with providing critical analyses of those reviews prior to, and during, the conference, as well as during the stakeholder meetings. The recommendations made in this paper were informed by the invited reviews, the discussions at the conference and stakeholder meetings, and finally, by the experience of the authors themselves. The resulting manuscript is not a policy document. Instead, it was aimed at providing recommendations and counsel for practitioners, scientists, and managerial staff.

Employment standards research has been of scientific interest for several decades (Davis et al. 1982; Gledhill and Jamnik 1992*b*; Jackson 1994; Payne and Harvey 2010; Shephard 1991; Shephard and Bonneau 2002; Sothmann et al. 2004; Tipton et al. 2013) and this is especially true in occupations with responsibility for public safety and security (e.g., law enforcement, structural firefighting, wildland firefighting, military). In theory, employment standards help to place, and to keep, the right people on the job while preventing putting the wrong people in harm's way. The financial, human, and property costs of incorrect employment decisions are substantial. Currently, there are few resources available to advance knowledge and support best practice in this field, despite the fact that the consequences of poor practice threaten employee health and organizational capability, while simultaneously placing a significant financial burden on employers and threatening the livelihood of otherwise capable workers.

Possibly because of the nature of organizations that historically develop and implement employment standards, a significant fraction of the research in this field has not been published in peer-reviewed journals. There are several problems with this practice. First, rigorous peer review invariably improves scientific work products. Second, publication is viewed by many as the final step in any scientific process, and in the absence of that step, the work may appear unfinished. Finally, in the event that an employment standard is challenged, peer-reviewed publication adds an important degree of credibility to the defense. One of the objectives of these dedicated conferences has been to raise the scientific profile of work in this field. In order for this objective to be realized, the quality of the work must meet the standards required by scientific journals.

Physiological readiness for work is usually measured using laboratory or field tests under standardized and controlled conditions. Test subjects are generally aware of test protocols and can prepare appropriately. While tests are often physically demanding and are sometimes accompanied by environmental stress, sensory deprivation, and anxiety-related stress, those stresses are minor compared with the extreme work conditions that emergency responders often encounter. Unfortunately, the discrepancy between relatively ideal test conditions and unpredictable work conditions has rarely, if ever, been addressed. An important challenge in this field will be to explore how physiological performance measured in the laboratory may be affected by factors that responders face in the workplace. Assuming that physiological capabilities, and hence work performances, are likely to deteriorate under real-world conditions, it would be prudent to account for those impairments from factors such as load carriage, environmental conditions, protective ensembles, hunger, dehydration, fatigue, and psychological stress.

It is not our aim to prescribe any one set of best practice procedures for international use; that would be presumptive and inappropriate at this time. Instead, the purpose of this paper, in combination with the accompanying reviews, is to begin to address ways through which this important field of study may be advanced towards more uniform and defensible practices. In the interest of helping to foster continued growth in the quality of scientific and professional practice in the field of employment standards, the following sections offer commentary on areas of concern, recommendations to improve scientific and professional practice, and where applicable, priorities for future research.

Delineating the field of study

Employment standards research and scholarship heavily emphasize the physical, physiological, and occupational medicine domains. Investigators identify and quantify physical stresses within essential elements of the working environment. Those stresses will often challenge homoeostasis, and by quantifying the regulatory responses induced, researchers can evaluate and understand the physiological strain experienced during work. Moreover, those responses help determine the attributes commensurate with stress tolerance, work capability, and injury resistance. This integrated field includes the collection and analysis of physical and physiological data, the development of relevant screening tests, articulation of the desired levels of physiological performance, identification of test scores that best define acceptable levels of achievement, and the development of professional practice guidelines. Therefore, employment standards research has a direct bearing upon jobs for which recruitment and retention are evaluated using screening tests.

Within this document, and the accompanying manuscripts, the emphasis is primarily upon the physical and physiological domains. In the past, the generic discipline name of *physical employment standards* has described research that involves human physiology and other disciplines such as psychology, ergonomics, and human factors. In some applications, the measurements of interest are physiological, while in others, a combination of physical and physiological variables is evaluated. In yet others, some interaction between physical burden and physiological ability may exist, with readiness for work inferred from a test completion time. The roots of this field are in work physiology but there have been many critical contributions from other disciplines. An essential developmental step will be to reconsider the name that best describes this multi-disciplinary field of research.

The correct implementation of physical and physiological employment standards should lead to the identification of individuals who are well suited to the demands of the workplace, who can meet job performance expectations, and who will sustain the capability of the workforce and its productivity. An equally important objective for the more physically demanding jobs is reducing the risk of workplace injuries (duty of care or due diligence). Many

such injuries are preventable, including those accompanying the use of overly protective clothing and equipment (Goldman 2001; McLellan and Havenith 2016; Taylor and Patterson 2016), arduous materials handling (Knapik et al. 2004; Taylor et al. 2015*a*, 2016), and ineffective screening procedures and standards that result in recruiting and retaining higher-risk individuals. Employers cannot abdicate their responsibilities to ensure a safe and healthy working environment.

A clarification of nomenclature

A critical step towards delineating a recognized scientific field involves the clear and consistent use of terms and phrases. To date, this has not been adequately addressed in the field of employment standards. It is recommended that less precise terms such as physical fitness be replaced with a more systematic phrase, such as readiness for work. The latter implies the requirement for physiological attributes consistent with objectively established demands of the workplace. Within the literature, terms such as standards and cut-scores have been used inconsistently and often synonymously. That practice does little to foster clarity in either scientific or professional practice. We therefore recommend adopting the systematic nomenclature outlined by Rogers et al. (2014) and further developed by Zumbo (2016), some of which are highlighted below.

Performance standards

Performance standards are defined as qualitative descriptions of the necessary attributes (e.g., knowledge, skills, competencies, behaviours) exhibited by individuals at distinct performance levels, with a clear delineation existing between adjacent performance levels (Rogers and Ricker 2006; Rogers et al. 2014). In the context of readiness for work, those standards describe levels of capability that distinguish between acceptable and unacceptable performance with respect to the safe, effective, and reliable execution of the essential job demands (Rogers et al. 2014). Those descriptions do not include the setting of numerical pass and fail scores (Kane 1994). Instead, standards are descriptive and qualitative, and they define work-related tasks, conditions (e.g., environmental factors, protective ensemble, physical burden), performance intensities, and, sometimes, a physiological reserve for operational safety. In this case, the notion of safety extends beyond personal needs, and includes co-workers and the public. Such prescriptions relate more to physiological attributes (e.g., endurance, strength, power) rather than to physical characteristics (e.g., height, mass, age), although they are not mutually exclusive. While some physical characteristics influence physiological function (Bowes et al. 2015; Notley et al. 2016; Schmidt-Nielsen 1984), standards based on those characteristics alone may be inappropriately discriminatory (Hogan and Quigley 1986).

Physiological aptitude tests

Tests should provide valid and reliable numerical evaluations (Milligan et al. 2016) of the physiological attributes of potential employees and incumbents. Those data must be directly relevant to evaluating achievement relative to the acceptable performance standards (Groeller et al. 2015), and thereby furnish an assessment of physiological readiness for work. Such tests are presently known as physical fitness, capacity, or aptitude tests, but those names also lack precision. Since such tests identify individuals with the physiological attributes consistent with specified levels of work performance under clearly articulated conditions, then it may be more meaningful if those tests were referred to as physiological aptitude tests. However, since aptitudes are not static phenomena, and can be acquired when absent, or enhanced and diminished when present, then an aptitude for work is at least partly under the control of the worker.

Cut-score

Cut-scores are points on the test scale used to delineate the levels of performance described in the performance standards, and to differentiate between acceptable and unacceptable performances (Rogers et al. 2014). The performances of individuals at, or below, a cut-score should be discernibly different from those above. In the case of a test designed to indicate readiness for work, the important distinction is whether the worker can perform critical duties at an acceptable level or work rate, and that score must faithfully represent, within acceptable limits of validity and reliability, that level of performance.

Minimal versus acceptable standards

The phrase minimal standard is extensively, but often loosely, used within occupational physiology, and it implies a level of achievement that has been deemed necessary (acceptable) to achieve a performance standard. The notion of satisfactory performance is widely variable. In many jurisdictions (e.g., Canada), legislation dictates that employers must consider hiring or retaining individuals who meet minimal standards of readiness. However, minimal often implies a low level of achievement, although in reality it should indicate that a reasonable and necessary expectation for safe and effective work performance has been satisfied. Importantly, the necessary performance threshold may represent quite high achievement levels for some physiological attributes. Therefore, to provide greater precision to the articulation of such thresholds, it is recommended that *minimal standard* be replaced with *acceptable standard*. While in theory, the two terms may be equivalent, *acceptable* has a far broader meaning, and is less ambiguous and demeaning.

Evaluating employment standards, physiological aptitude tests, and cut-scores

True and false positives

The effectiveness of existing employment standards and screening procedures can be evaluated by determining the probability of correctly differentiating between individuals who can meet these acceptable expectations, and those for whom the job is too physically demanding. In this context, high-resolution discrimination is desirable, and indices of discrimination permit the accuracy of these practices to be evaluated (Schulzer 1994; Swets 1986). This seems paradoxical, for much energy is invested into processes that minimize other forms of workplace discrimination (Adams 2016; Hogan and Quigley 1986). However, an inability to differentiate between high- and low-risk workers can be as legally burdensome as workplace discrimination because of the failure to fulfil reasonable duty-of-care obligations, so high-resolution procedures are desirable from both capability and injury-prevention perspectives.

Appropriately identified individuals are known as true positives, while correctly eliminated candidates are the true negatives. Imperfections within these processes may lead to errors of discrimination when potentially capable workers are not selected or retained (false negatives or rejections), and when less-capable individuals are inadvertently recruited or retained (false positives or selections). In the first instance, there has been a missed opportunity, and in the second there is an elevated probability of exposing less-capable workers to high-risk conditions.

The ratio of true positives to the sum of the true positives and false negatives (expressed as a percentage) defines screening sensitivity, or the possibility that ideal candidates will be correctly identified through those procedures. Similarly, procedural specificity (correctly identifying less-capable individuals) can be derived as the ratio of true negatives to the total number of true negatives and false positives. It is often impractical, and sometimes impossible to determine either the true or false negative outcomes. Nevertheless, differentiating between true and false positives may be feasible, and it is recommended that time be invested into

estimating the percentage of capable workers identified relative to all who satisfied a cut-score; the positive predictive index.

$$\text{test sensitivity} = \text{true positives}/(\text{true positives} + \text{false negatives}) \times 100$$

$$\text{test specificity} = \text{true negatives}/(\text{true negatives} + \text{false positives}) \times 100$$

$$\text{negative predictive index} = \text{true negatives}/(\text{true negatives} + \text{false negatives}) \times 100$$

$$\text{positive predictive index} = \text{true positives}/(\text{true positives} + \text{false positives}) \times 100$$

The separation of true and false positive outcomes often occurs beyond recruitment, and such derivations may be beyond the scope of the investigative team that developed the screening tests, derived the employment standard, and set the cut-scores. However, this does not mean those indices should be ignored. We suggest that the obligation for undertaking these computations falls to the work organization, for that group alone has the capacity to evaluate work capability and to identify the false positive workers. Furthermore, evidence such as the positive predictive index will be very useful when dealing with legal challenges (Adams 2016; Hogan and Quigley 1986). It is therefore recommended that employers should, as part of their on-going employment standards review, consider determining their own positive predictive index, although this may represent a legally difficult area in some jurisdictions. For this, they already have the number of positive outcomes; those who satisfied the employment standard and its corresponding cut-score. All that is required is to differentiate between the true and false positives (i.e., those who, following selection, were found not to have the required physiological attributes) without bias.

Cut-score uncertainty

It must be recognized that employment standard absolutes are difficult to justify, and a degree of uncertainty exists around every cut-score. In fact, researchers must always be prepared to answer the question: how certain are you that an individual with this score can do the job, while another person with a very similar, but slightly lower, score cannot? In almost every situation, the answer must contain some uncertainty, and the implication of this is that zones of uncertainty need to be considered and incorporated into cut-scores (Fullagar et al. 2015; Rogers et al. 2014; Tipton et al. 2013).

Nevertheless, investigators have not always acknowledged this uncertainty, the result of which is that many cut-scores have been considered as absolutely precise. Examples of allowing for this uncertainty may be found in Rogers et al. (2014), where the recommended cut-score was adjusted to account for disagreement among the 25 judges whose workplace expertise was utilized to set a cut-score, and also in Fullagar et al. (2015). This uncertainty influences the identification of true positives and negatives when test results fall near the cut-score.

In cases where cut-scores have been based on the normative analysis of performance scores of a workforce sample (e.g., Gledhill and Jamnik 1992*b*, 2011), there is the potential for introducing bias if the sample size is small or unrepresentative. If such bias exists then the cut-score might be more reflective of the fitness characteristics of the small sample rather than the fitness requirements for performing the job safely and effectively. It is recommended that deliberate sampling strategies be employed to ensure that large samples of workers are evaluated to accurately represent the complete range of the workforce in question with regard to characteristics such as age, sex, size, experience, physical fitness, and capability (Jamnik et al. 2010*b*; Sothmann et al. 2004). Cooperation is required from the workplace stakeholders (e.g., management, union) to understand the true nature of the workforce, and to identify and recruit individuals possessing those characteristics.

A common format for work simulation tests involves a test circuit, where the outcome is based on the completion time of a series of work-related tasks. Ignoring measurement errors and motivation, performance variability is generally related to test familiarity and biological variability. The former represents systematic variability and the latter, random variations. Boyd et al. (2015) concluded that, for most individuals, at least three practice trials were necessary before test scores attained a fitness-dependent asymptote. Similarly, Fullagar et al. (2015) reported the elevation of some subthreshold test performances following test familiarization. Thus, without adequate practise, it may be impossible to interpret scores close to the cut-score that result from this type of test. This should not be interpreted as a criticism of the circuit-test format, but rather a caution concerning the appropriate use of such tests to evaluate physiological status.

Knowledge of the extent that biological variability may influence performance allows an organization to develop strategies to address test scores that fall within that range. Several approaches may be considered: granting a pass to those whose test scores fall below the cut-score, but within the range of biological variability; re-testing with the expectation of meeting the cut-score; or ruling that all must be able to achieve the cut-score on their worst day. These diverse examples demonstrate that organizations must develop a defensible policy relating to the handling of test scores falling within this category of variability.

Creating and implementing employment standards: the process of research

In the realm of occupational physiology, much attention has been applied to procedures for establishing employment standards that will withstand close examination in court. Following clear, logical, and well-considered methods should lead to fair and equitable employment standards that serve the intended purpose and be most defensible. In this section, the necessary research phases are briefly highlighted with a view to facilitating the development of appropriate and legally defensible employment standards and cut-scores for physically demanding jobs (Adams 2016; Hogan and Quigley 1986; Zumbo 2016).

A sequential, procedural framework for the important aspects of employment standards research, arising from previous research, is presented in Table 1. Those 20 steps are aimed at selecting and retaining a capable workforce (true positives), while simultaneously reducing workplace injuries, inappropriate discrimination, and adverse impacts upon employers. Ideally, that research will yield procedures and standards with a high positive predictive outcome. A comprehensive treatment of the procedures was assembled by Gledhill and Bonneau (2001), and the current evolution reflects contributions from Shephard (1991), Constable and Palmer (2000), Taylor and Groeller (2003), Payne and Harvey (2010), and Jamnik et al. (2013).

It might be argued that most steps in Table 1 represent critical points. Nevertheless, unless step 2 results in the identification and involvement of the correct people (the management team), then it is likely that potentially incorrect decisions may occur during many of the subsequent processes. This team will evaluate and approve progress on at least five occasions (Table 1: steps 6, 10, 13, 15, 17), and should include individuals found in academia, government organizations, and the private sector who were selected on the basis of three different attributes. First, knowledgeable and experienced employees from the target workforce are needed to help the scientists learn about and understand the job while providing easy access to the necessary resources. These individuals are the link between the job and the scientists. The second group

Table 1. Procedural framework for developing employment standards and cut-scores for physically demanding jobs (Gledhill and Bonneau 2001; Taylor et al. 2015*b*).

Phase	Step	Description
1	1	Justify establishing an employment standard
	2	Appoint a management team with appropriate knowledge and experience
2	3	Familiarize the research team with job requirements and duties
	4	Preliminary job review and analysis
	5	Identify the essential and physically demanding tasks
	6	Approve and validate the list of essential and demanding tasks
	7	Produce a subset of tasks using employee surveys or focus groups
3	8	Characterize those tasks: observe, measure, quantify
	9	Identify the criterion tasks
	10	Approve and validate the criterion tasks
4	11	Develop physiological screening tests
	12	Standardize screening tests, including administrative procedures
	13	Approve and validate screening tests and procedures
5	14	Evaluate screening test validity and reliability
	15	Approve standard development for test performances
6	16	Develop test performance standards and cut-scores
	17	Approve and validate test performance cut-scores
	18	Implement screening test(s)
7	19	Develop instructional and preparatory guidelines for candidates
	20	Review the screening process and outcomes as the job changes

Note: From Taylor et al. (2015*b*), modified with permission of J. Occup. Environ. Med., Vol. 57, p. 1064, © 2015 Wolters Kluwer Health Inc.

will almost invariably come with a sound and demonstrable understanding of the scientific method (e.g., peer-reviewed publications). Although familiarity with the job can be relevant, it can also be an impediment if it restricts vision. The final members will be senior administrators within the organization who must also be able to put aside potential areas of bias established prior to entering this role. None of these groups must dominate the composition of the team, with balance being the key. Consideration should be given to adding personnel from human resources with human rights and legal expertise.

The issues of sex bias must be considered. Historically, many organizations involved with employment standards have been male-dominated (e.g., law enforcement, firefighting, military). Consequently, there may be unintentional bias in factors such as equipment handling, training, and operational protocols (Friedl 2016). From the outset, all team members should be aware of this possibility and consider alternative solutions.

The second research phase must lead to identifying the essential, physically and physiologically demanding tasks within the occupation. Although there are many ways to arrive at that task list (e.g., Anderson et al. 2002; Blacker et al. 2015; Blacklock et al. 2015; Jamnik et al. 2010*a*; Rayson et al. 2009; Reilly et al. 2006*a*; Taylor and Groeller 2003; Taylor et al. 2015*b*), due consideration must be afforded to the importance, difficulty, frequency of performance, and duration of those tasks. For instance, while an immensely difficult, but rarely performed, activity might not represent the typical working environment, failure to successfully complete that task may have catastrophic operational outcomes. An example of how one might work through those challenges was presented by Taylor et al. (2015*b*). In decisions such as these, however, informed subjectivity is frequently involved (Tipton et al. 2013).

In the third research phase, the physical and physiological demands of the job are determined (e.g., Gledhill and Jamnik 1992*a*; Jamnik et al. 2010*b*; Reilly et al. 2006*b*; Taylor et al. 2015*c*). Such demands, which will form the basis of subsequent employment standards, should be initially categorized through real-time observation. Subsequently, they should be evaluated using a representative and sufficiently large workforce sample when performing realistic workplace simulations at operationally appropriate intensities. When those activities involve the impact of loads, either in the form of personal protective clothing and equipment or carried masses, the contemporary scientific consensus is that the metabolic costs of work should be reported in absolute units (McLellan and Havenith 2016; Taylor et al. 2016), and not as mass-specific equivalents (Royal Society (Great Britain), Symbols Committee 1975). The outcome from this work will be a list of criterion tasks that best represent the diversity of tasks that may be encountered in the workplace, and there are several ways to arrive at that list (e.g., Taylor et al. 2015*c*).

Although a consensus exists on the broad methodological approach used within the first three research phases, the transition between criterion task identification and the development of representative, valid, and reliable screening tests (phases 4 and 5) has not been well described. Therefore, this process was recently elaborated by Groeller et al. (2015), with the objective being to ensure that individual test items provided reliable appraisals of the criterion-task performances and physiological aptitudes. When assembled into a test battery, the outcome must be a screening test possessing both construct and criterion-related validity, and it must assess the full breadth of physiological attributes necessary for working with an appropriate level of competence, and without undue risk of injury.

The penultimate phase of this process involves the development of test performance standards and cut-scores. Researchers have followed a number of methods to arrive at effective selection and retention outcomes (Blacker et al. 2015; Blacklock et al. 2015; Fullagar et al. 2015; Jamnik et al. 2010*b*; Reilly et al. 2006*b*; Rogers et al. 2014). Those outcomes could eventually be subjected to legal scrutiny (Adams 2016; Hogan and Quigley, 1986). Therefore, several helpful and relevant topics are described in the accompanying review articles (Milligan et al. 2016; Zumbo 2016).

Finally, the requirement to re-visit that developmental process at regular intervals, either partially or in detail should be anticipated. This is particularly important when operational procedures are modified, when new equipment is deployed, and when new roles are added to the job description. These iterative processes must not, however, impose an unreasonable hardship (adverse impact) upon the employer (Hatfield 2005). Organizations should recognize the importance of this step at the time of implementation. In some cases, progressively more precise information on test and cut-score validity can gradually be acquired after implementation, including an evaluation of the predictive capacity of those tests. Therefore, ongoing analysis of test and

cut-score performance by the organization, in cooperation with researchers, is highly recommended.

This section closes with a recommendation that employment standard-setting methodology and procedures be incorporated into an international standard. This step may bring the widely varying approaches that are now commonly used within this discipline closer to a common set of internationally accepted practices. Such a step would also have the advantage of making different projects more comparable, while helping to bring a level of quality assurance to research and practice in this field. However, there must still be room for individual variations within these processes, for only with such freedom can innovation and advancement occur. It is therefore recommended that the organizers of future conferences facilitate discussion on the development of such an international standard.

Considerations beyond standard development

A comment on recycling

It is natural to expect, if financial savings can be made, that managers would prefer to adopt tests developed for other, similar organizations. Unfortunately, there are no off-the-shelf screening tests, standards, and cut-scores for any single occupation. The only circumstance where this may become appropriate is when exactly the same tasks are performed using identical equipment, clothing, and techniques; when those tasks are performed at the same intensity, frequency, and duration; when the acceptable standard of workplace performance remains the same; and when the original test has been demonstrated to be a valid predictor of work-related performance within the original job. When one or more of those states do not coexist, then the legal defensibility of recycling is open to interrogation (Hogan and Quigley 1986). Therefore, previously established screening tests and employment standards need to be intermittently revisited, both within the target organization itself, and among organizations with similar job descriptions, regardless of the validity of those tests, standards, and cut-scores when originally determined (e.g., Bonneau 1996). This need not necessarily involve every step outlined in Table 1, although components from most phases will require consideration. Accordingly, it would be imprudent for organizations to recycle screening tests and employment standards developed for others without carefully considering the implications of that practice.

Employment screening test frequency

The frequency of screening test administration for incumbent workers is often debated. One position is that when testing only occurs at the point of workforce entry, employers could be guilty of failing in their duty-of-care obligations. In some jobs, critical incidents happen intermittently and can be separated by long periods of lower intensity work. This duty cycle can be offset with regular training for critical incident responses, but unless carefully monitored, there is no guarantee that training will maintain physiological readiness. Some workers will spend their entire career within such cyclical variations in operational tempo. Unless those workers are habitually active in areas that complement the physical demands of the workplace, operational capability will almost certainly decline, and the risk of workplace injury will rise (Kenny et al. 2016; Storer et al. 2014). One solution is the more frequent administration of employment screening tests. In this way, the employer can identify workers requiring assistance with maintaining health and work-related fitness.

Several counter arguments must also be recognized. Frequent testing can be costly and reduce operational capability during test periods. The testing of incumbents can appear to threaten job security. It can be hard to determine the appropriate time gaps between tests. Since some workers are diligent in their commitment to health and fitness maintenance, testing might appear as a lack of trust within the workforce. These are significant matters, but they are not insurmountable, providing all parties can be brought to a common agreement on the merits of repeated testing. Just as there can be no off-the-shelf screening tests, there is no single solution to this dilemma. However, it is recommended that test frequency be geared to the age of each worker, as the decline in physical and physiological capabilities within typical adults is time dependent (Kenny et al. 2016). Those changes are also nonlinear (Groeller 2008; Proctor and Joyner 2008), with the rate of degradation increasing within and beyond the sixth decade, so testing frequency must also reflect the typical age-dependent decline in physiological function.

The demands of some occupations could require annual testing throughout the career. At the very least, it is recommended that the re-testing of incumbents should occur after 10 years of employment or at the age of 40 years, whichever occurs first. Any age to signify the commencement of re-testing might be viewed as arbitrary, although the age of 40 years has historical precedent in the identification of older workers (US Equal Employment Opportunity Commission 1967). During the fifth decade, it is suggested that testing be administered at the ages of 45 and 50 years. During the sixth decade, it is recommended that testing be more frequent, and at intervals of no greater than 2 or 3 years. Legislation in many countries specifies the responsibility of the employer for duty-of-care, and this testing may be considered part of that responsibility. Employees in physically demanding occupations, especially those that can threaten the safety of co-workers and the public, must also accept a personal responsibility to maintain physiological readiness for work.

A frequently encountered argument against regular physiological testing is that aging workers are promoted to supervisory positions with reduced physical demands, making testing for front-line work unnecessary. While this can be true, it would be erroneous to assume that all older employees fit that model. We suggest that such employees will benefit from a more health-focussed evaluation, especially in occupations such as firefighting where it is well documented that cardiovascular incidents are the main cause of on-duty deaths (National Fire Protection Association (NFPA) 2014). We further suggest that for older employees who regularly face front-line duty in such occupations, advancing age (increased health risk and reduced functional capacity) makes more frequent evaluation of health status and physiological readiness more important than ever. In summary, organizations should recognize that rank and age should be considered when deciding the focus and timing of physiological evaluations.

A subtopic of this section concerns the return to work of employees who have been away from a physically demanding job for a significant duration, through illness, injury, or redeployment. This is a complex topic, as it includes individuals who were bedridden for a significant duration, workers performing equally demanding work elsewhere, and those reassigned to administrative duties. There is a natural, and perhaps justifiable, temptation to view each group as a different subset. Yet there is clear research evidence that many physiological similarities exist between the first and last groups (Narici et al. 2008). Indeed, habitual sedentary behaviours come with physiological penalties (Chakravarthy 2008), and these can impact upon readiness for work and work-related injuries (Poplin et al. 2014; Storer et al. 2014). To address the issue of returning to work within physically demanding trades, it is the authors' recommendation that two prerequisites must be satisfied in sequence. First, the worker must be evaluated by an appropriately trained, medically qualified specialist regarding the risks associated with performing the full battery of health and work-readiness screening tests. Second, the worker should be required to undertake those tests and successfully attain the acceptable standard and cut-scores for that job. Clearly, these steps call for a better integration of medical support into the evaluation of readiness for return to work.

Adverse impact on employers

Recycling employment standards may be appealing to avoid the time and expense of developing organization-specific screening tests. However, if recycling is not feasible, such arguments should be outweighed by the benefits that accrue to the workers, such as general health, quality of life, and reduced injures (Booth et al. 2000; Powell and Blair 1994), and to the organization, such as lower legal expenses (Adams 2016), reduced sick leave (Proper et al. 2006), and fewer workplace injuries (Taylor and Taylor 2012). Nevertheless, there must be a threshold below which the cost–benefit ratio becomes disproportionately large and unreasonable. In these situations, it is the employer's obligation to demonstrate that the burden of establishing employment standards is too great, and represents an adverse impact (Hatfield 2005).

Another area of concern for employers, also within the realm of adverse impact, is the consequence of employing and accommodating workers with special needs. In occupations such as firefighting, for example, the difficulty and the impact of accommodation will increase with the level of disability. As well, the relative impact on safety and effectiveness (both of the involved worker and co-workers) must be evaluated on a case-by-case basis. Much can be achieved through the application of common sense. When circumstances change for a worker, however, possibly because of aging, illness, or injury, employment within an organization may be guaranteed by good will, policy, or contract. However, continued placement or return to work within specific physically demanding trades within that organization may not be feasible. Outcomes from appropriately developed, physiologically and medically based employment standards research will often provide independent evidence that can support reasonable decisions. Another question that arises is whether or not equipment and operational practices can be modified to allow workers with special needs to satisfy those standards; these issues are discussed in the following section. In Canada, for example, employers are expected to provide accommodation up to the point of undue hardship (Supreme Court of Canada 1999). While the term undue hardship is not well defined, the message is clear that employers are expected to try to accommodate workers with certain characteristics (see Eid 2001). Valid and reliable employment standards are valuable tools to help navigate these difficult areas.

Accommodating individual differences

Considerations regarding the accommodation of individual differences must, within reason, represent normal recruitment and retention policies (Hatfield 2005; Hogan and Quigley 1986), yet some employment practices will result in the exclusion of particular individuals. If those practices are inappropriately discriminatory, then it may be argued that they have an adverse impact upon those members of society. Thus, adverse impact must be viewed through two equally powerful lenses; one that is focussed on the impact on the employer, and another that is used to assess impact upon the individual.

In some working environments, the modification of equipment and operational practices can be used to allow people with limited capabilities or special needs to satisfy the acceptable employment standards and cut-scores. Employers must give such measures reasonable consideration. However, employers vary in their capacity for adopting forms of accommodation, with some special needs unable to be accommodated by any employer.

In accompanying articles (Roberts et al. 2016; Taylor et al. 2016), the relative impact of absolute (standardized) loads on workers of varying body size is discussed in detail. The reality is that smaller individuals must work relatively harder to carry the same load (e.g., mission-essential equipment, tools, weapons, protective clothing) as their physically larger counterparts. Yet research clearly shows that size alone does not adequately predict work performance under load-bearing conditions (Phillips et al. 2016). Many of the physiological underpinnings of work performance can be altered through exercise training, with improved physiological function often counteracting differences in physical characteristics (e.g., age, sex, mass, stature). It is strongly recommended that potential and current employees take personal responsibility for improving their physiological readiness for work. By the same token, employers are encouraged to provide appropriate support (e.g., exercise facilities, time to train) as a form of accommodation.

On the use and misuse of analytical procedures

In accompanying communications (Milligan et al. 2016; Zumbo 2016), details are provided concerning the appropriate evaluation of measurement reliability, reproducibility, and validity. These are backbone procedures of sound scientific practice. In addition, there is a wide range of ancillary analytical tools, and these too must be applied correctly and with a sound understanding of their strengths and limitations. While it is beyond the scope of this contribution to comprehensively address the full breadth of relevant topics, we have chosen to highlight selected points that would seem to require attention. Within the context of employment standards research, statistical imprecision and misuse can have a direct bearing on employee selection, on perceived or real discrimination within that process, on the incorrect recruitment of less capable individuals (false positives) and elevating the risk of workplace injuries, and, on the legal defensibility of employment standards and cut-scores.

Greater uniformity in reporting experimental observations in this field of study is recommended. With regard to raw data normalization relative to body mass, we must always be alert to the possibility that such practices may be invalid, potentially leading to the creation of bias within those data (Packard and Boardman 1999). This is discussed in detail in an accompanying manuscript (Taylor et al. 2016), and is reinforced below. The case is presented that arithmetic (linear) mass normalization should not be undertaken unless the raw data pass through zero, or have been adjusted to do so. Next we need to consider which information concerning the experimental observations is most valuable to readers, in addition to the means. We have several options: standard deviations, standard errors, confidence intervals, variance, and data ranges. Because employment standards investigators are interested in describing and understanding data distributions, it is recommended that standard deviations be reported, accompanied by data ranges. Nevertheless, because the other parameters can be derived from the standard deviation, authors are also requested to provide sample sizes.

Correlation coefficients and the coefficient of determination

In circumstances where one may wish to predict changes in a dependent variable, or to assign a causal relationship between independent and dependent variables, it is a common practice to evaluate the goodness-of-fit, and thereby derive a correlation coefficient (r). This is most commonly performed where such relationships are linearly associated, and using the least-squares, best-fit approach to minimize the sum of the squared residuals. When the correlation coefficient is squared (r^2; coefficient of determination), one obtains an estimation of the extent to which variability within the dependent variable may be explained on the basis of variations observed within the independent variable. Of course, there is no proof of causality, just a quantification of association, and a statistically significant relationship does not imply physiological significance or importance. Moreover, correlations do not quantify the level of agreement between those variables (Altman and Bland 1983).

When undertaking regression analyses, there are procedural considerations that are covered in standard texts. Herein, emphasis is on the number of observations and the nature of the data used to perform regression analysis. There is no fixed number of required data points, although we know that the more points we

have, and the wider and more even their distribution along the relationship in question, the more confidence we can have when attempting to interpret correlations. Data clustering is problematic, as are outliers and limited data points at either end of the distribution. Much can be achieved through visual examination. In addition, statistical power will also be influenced by the sample size and the effect size. Of course, the required sample size for an optimally designed experiment varies across disciplines, and across measurements within disciplines. Within occupational physiology, where inter-individual variability can often be significant, it is recommended that sample size be determined based on recognition of the need to adequately reflect the population of interest. The likelihood of error decreases as sample size increases. Evidence of power calculations should be included when possible. To adequately represent the diversity of the workforce, larger samples will be required. For example, if one were to stratify the workforce by categories (e.g., age by decade), it would be necessary to achieve at least the minimum sample size within each category.

When looking for differences between experimental treatments, one evaluates whether or not the outcome means were statistically significant. When undertaking regression analysis, one must not use those means, but individual data points. That is, every point contains information related to the within-subject (biological) variability of the dependent and independent variables. Moreover, sequences of points from different individuals inform us about between-subject variations. Therefore, to faithfully describe both of those relationships, and to avoid an artificial inflation of the correlation coefficient, regression analysis must be performed using individual data and, whenever possible, that should involve regressions on each participant, with the resulting coefficients determined as averages (de Rey et al. 2001; Notley et al. 2014; Taylor et al. 2012*b*).

Bland–Altman plots

The Bland–Altman method (Bland and Altman 1986) is often used to evaluate whether or not 2 measurement techniques may provide the same information, and might therefore be used interchangeably. The primary outcome is the level of agreement between the methods. The focus is not just upon the correlation between methods, although that is important, but on their difference. This approach was first developed for appraising different methods for quantifying clinically relevant physiological functions (Altman and Bland 1983), but the method has many applications, including employment standards research (Notley et al. 2015).

The method is based on determining the difference between values derived from a criterion measurement of the variable of interest (value A) and a second, perhaps predictive, derivation of the same variable (value B). The methodological bias (A – B) is then plotted against the combined mean of those measurements ((A + B)/2) across a physiologically relevant range. Such a plot includes information related to reliability, validity and measurement error, and bears some resemblance to a plot of residuals. Agreement limits between the methods under evaluation are then defined using 95% confidence intervals relative to the mean bias. Unfortunately, one often finds that the most critical aspect of this analysis has not been considered. That is, such plots can only be meaningful if one has an *a priori* expectation of what may be an acceptable limit of agreement, and against which the observed deviations might be compared (Bland and Altman 1999). While the method quantifies agreement between 2 measurements, it is the responsibility of the investigator to determine whether or not those agreement limits are acceptable and meaningful (Giavarina 2015). Unless this essential step has been completed, the current authors would argue that the outcome of such an analysis is potentially invalid.

To illustrate the difference between agreement limits based upon statistical derivations and those based on acceptable limits that experienced scientists might deem to be appropriate, consider Notley et al. (2015). That group evaluated the utility of predicting the oxygen cost (criterion variable) of firefighting activities from indirect physiological measurements (heart rate and minute ventilation: predictive variables). Both indirect variables are independently and strongly correlated with the criterion variable (Booyens and Hervey 1960; Durnin and Edwards 1955), but can they be used interchangeably with that variable, and will they yield similar outcomes in the field? To arrive at an acceptance threshold of relevance to occupational physiologists, Notley et al. (2015) invited scientists from 8 countries to define a meaningful oxygen consumption threshold for evaluating the predictive precision of those 2 variables. The threshold for the acceptable limits of agreement was then independently set by 25 scientists (14 universities, 11 government institutes: 382 years of experience in occupational physiology): ±0.24 $L \cdot min^{-1}$. When that threshold was applied, both surrogates were found to overestimate the oxygen consumption of those field tasks to levels that exceeded the measurement precision required by experienced physiologists. That is, neither surrogate was acceptable as a valid and reliable predictor of the oxygen cost of those workplace activities. Had the acceptable limits of agreement not been appropriately defined, then the authors may well have produced statistical support for both surrogates that would mislead subsequent investigators into believing that, within such an occupational setting, one might use either index instead of taking direct measures of oxygen consumption. That interpretation has been made by others, yet it seems quite indefensible.

Receiver operating characteristic curves

A method for comparing criterion and predictive variables, and one that is becoming increasingly common within physiology is the derivation of receiver operating characteristic curves (Hanley and McNeil 1982; Metz 1978). This technique has an interesting history, and has become favoured for evaluating the outcomes of binary classification systems, such as those used for the classification of potential recruits, and predictive power of those procedures. For example, it is often used to evaluate the diagnostic precision (discriminatory utility) of clinical procedures (Schulzer 1994; Swets 1986), with that outcome derived from the graphical relationship between the false positive identification rate (abscissa) against the rate of true positive identification (measurement sensitivity). A numerical evaluation is derived from the area under the curve of that relationship, the magnitude of which is used to quantify predictive power (1 = perfect prediction, 0.5 = random prediction).

The development of sensitive and highly specific employment screening tests with good predictive power is also a primary emphasis of employment standards research, and some may consider this analytical approach to also be appropriate within that context. However, the method is not without limitations (Halligan et al. 2015; Vanagas 2004), and the question of interest is whether or not the method can, or should be, applied outside the clinical environment. Recently, Taylor et al. (2012*b*) compared the utility of changes in saliva osmolality for detecting physiologically (3%) and clinically interesting (6%) instances of dehydration, using 2 analytical approaches. The resulting receiver operating characteristic curves indicated that saliva osmolality should have been a good predictor of hydration state. Yet, when those same data were analyzed using a double-threshold detection method (3% and 6%), <50% of the data were found to correctly identify individuals who were >3% dehydrated. Within the 3%–6% dehydration range, the sensitivity of the osmolality method was 64%, but when the identification of dehydrated individuals was most critical (>6% dehydrated), its sensitivity declined to 42%. Clearly, in these circumstances, these analytical approaches yielded divergent outcomes, leading one to question the usefulness of applying receiver operating characteristic curves to that question. Moreover, it pro-

vides another example of how such procedures need to be used with a clear understanding their limitations, and with a realistic appreciation of the physiological significance of the resulting outcomes. It is recommended that occupational physiologists resist the temptation to blindly follow analytical methods from other disciplines.

Systematic reviews with meta-analyses

The final topic in this section relates to levels of experimental evidence. While experienced researchers rarely require advice on how to rate the merit and significance of different types of research evidence, some granting agencies think otherwise, and this has led to the generation of merit tiers. At the top of the merit table, one invariably finds the systematic review. This has largely been driven by medical funding agencies, for this type of review (as opposed to the narrative form) lends itself to the extraction of useful interpretative information from vast collections of clinical evidence (Mulrow 1994). The analytical process involves identifying relevant research using explicit search strategies, including (or excluding) investigations from that pool according to well-defined criteria, and then consistently evaluating that work against appropriate methodological standards. This all sounds perfectly reasonable; but is it always reasonable?

Readers are encouraged to enter the following term into a PubMed search: *systematic review meta-analysis*. Now select a journal, such as *Sports Medicine*, that specializes in applied physiology reviews and see how many systematic reviews have been published. Examine some of those reviews to determine how many experimental subjects constituted the overall sample size for the meta-analysis. Now consider this; when performing a meta-analysis, it is standard practice to exclude under-powered studies (small sample sizes), yet Turner et al. (2013) have shown that "underpowered studies made up the entirety of the evidence in most Cochrane reviews". Look again at the systematic reviews in the realm of applied physiology. How many have included underpowered studies? Indeed, in how many reviews was the total sample size less than 100? From a clinical perspective, some would posit that samples sizes <1000 individuals fail to provide a useful patient sample (Thorlund et al. 2010). We must therefore evaluate whether or not each systematic review provides a reasonable and meritorious collection of strong scientific evidence. In their presentation of the case that small clinical reviews are not wasteful, Handoll and Langhorne (2015) concluded that when systematic reviews contained inadequately powerful studies, "the evidence was insufficient to inform practice." They then highlighted how the gradual development of a series of thematic reviews could subsequently inform practice. This is not questioned, but what must be more closely examined is the merit of the one-off, lower-powered systematic review; such reviews primarily highlight evidence deficiencies (Handoll and Langhorne 2015).

Accordingly, the nonclinical application of systematic reviews is cautioned. Indeed, the method is becoming overused, and this appears to be happening in areas for which that approach has insufficient power. In addition, some investigators seem to be erroneously assigning greater merit to this type of scholarship than to the intellectual evaluation of the experimental methods, and subsequent data interpretation that can be provided by experienced scientists (narrative reviews). Accordingly, the use of systematic reviews within occupational physiology seems ill-advised, unless a critical mass of well-designed and suitably powered primary investigations exists.

Highlighted topics

The accompanying reviews within this special issue emphasize critical discipline-related topics targeted for in-depth discussion, and readers are encouraged to consult those contributions. Below, we have extracted some of the more salient features, with supplementary commentaries, to highlight some decisive issues.

Age-dependency of employment standards

There is historical evidence for employment standards that differed according to age, although this is not the position recommended by the current authors. This approach could only be considered with clear evidence that work demands are reduced in parallel with advancing age. In the absence of such evidence, sliding standards are likely to be discriminatory, this time to the disadvantage of younger individuals. The greater concern is that operationally relevant requirements and acceptable standards for the successful performance of most jobs remain static and age-independent. To avoid aging workforces that might progressively become incapable of performing the required job, employment standards must remain age-neutral (or age-free). This is a principal tenet of the discipline.

Physiological (biological) variability exists among younger adults, but within older individuals, it becomes much more pronounced, resulting from lifestyle choices, chronic exposure to potentially hazardous workplaces and substances, injuries, and the presence of disease (Groeller 2008; Kenny et al. 2016). Furthermore, the literature on aging can present a confusing, and sometimes imprecise, overview because of the very powerful interactions of habitual physical activity and inactivity with physiological function (Chakravarthy 2008; van der Ploeg and Bauman 2008). Indeed, because even apparently healthy, chronically sedentary older adults may have subclinical pathological conditions (Blair et al. 1996; Jenkins and Plasqui 2008), aging is perhaps best studied in habitually active individuals who more closely resemble our evolutionary phenotype (Holloszy and Kohrt 2011; Pollock et al. 2015), as per the recommendations concerning therapeutic and intervention strategies (Booth and Lees 2006). In fact, the effects of aging and habitual inactivity may be additive. As described by Kenny et al. (2016), this interaction means that we cannot assume physiological status can be evaluated on the basis of chronological age alone. We must also consider physiological as well as functional age (ability to do the job).

While physiological aging is a plastic phenomenon, its consequences are not preventable (Proctor and Joyner 2008). Habitual exercise and healthy lifestyle choices may slow the rate of decline, or alternately, preserve higher levels of physiological function (Nelson et al. 2010; Pollock et al. 2015). Consequently, older workers who have satisfied the required employment standards, and particularly those who are less physically active, will generally be positioned closer to the cut-score than their younger co-workers. Without sound exercise habits that regularly and adequately engage work-related physiological attributes, those individuals will, sooner or later, enter the zone of performance uncertainty, and eventually the region of unacceptable performance. Notwithstanding the personal and social implications of that outcome, the employer stands to lose an experienced and valuable member of the workforce. It therefore becomes a shared problem, with both the worker and the employer participating in its resolution. Walker et al. (2014) suggested a strategy of setting higher entry thresholds, more frequent assessments, and on-going fitness for work programming for older firefighters. Inflated entry requirements, for the purpose of attenuating the effect of aging on work performance, would be very contentious in some jurisdictions, and cannot be recommended. However, the latter recommendations may well alleviate some of the concerns with aging employees.

Of course, if physiological aptitude tests are valid reflections of the work demands, then they will also be functional in nature. This means that some older workers, perhaps having pronounced physiological decrements, may actually demonstrate a younger functional age (work-related ability; Ilmarinen 2001). Thus, chronological, physiological, and functional ages must be viewed as potentially independent phenomena. Moreover, the reverse is also true, with some chronologically and physiologically younger workers possessing a much lower functional capacity.

Sex issues in employment standards

In some circumstances (e.g., women in the combat roles), equality of participation opportunity is a legislative matter that is beyond the control of the employer. While there is historical evidence for the application of different employment standards for men and women within the same job (Stevenson et al. 1992), ironically, those standards have usually been lower for women, and often have been linked to population norms rather than to objectively documented occupational requirements. However, a fundamental tenet is that sex-neutrality must be sustained when the standard is linked to physical demands that are common to all workers.

The anthropometric and physiological differences between males and females have been summarized elsewhere (Epstein et al. 2013; Roberts et al. 2016; Shephard and Bonneau 2002). The reference woman, for example, is smaller in stature and has less lean body mass than her male counterpart, and these characteristics frequently carry over to other determinants of physiological output, such as the cardiopulmonary system. These differences do not mean that females cannot do the same work as males. However, it should be recognized that females close to the characteristics of the reference woman, just like smaller males, must work at a relatively higher fraction of their peak physiological attributes (e.g., strength, aerobic power) when doing the same work as the reference male. Assuming that the force application and energy demands of a task are sex-neutral, the logic of lower physical fitness for work standards for females is untenable.

Earlier, the importance of striking the correct balance within the project management team was discussed. That principle carries forward to the completion of each step within phases 2 and 3 of the recommended procedural framework (Table 1), recognizing the need to consider that task completion by males who have been trained by males may reflect a bias that is likely unintentional. In many of those occupations (e.g., specialized trades in the military, structural firefighting), the lack of female workers presents real problems to researchers attempting to avoid bias. When undertaking research in historically male-dominated occupations, it is essential to avoid bias that may arise simply as a function of the organizational history. To better appreciate this problem, one might consider reversing the sex of the incumbents, and integrating males into a predominately female workforce (Friedl 2016).

The same potential for bias can arise when determining performance standards and setting cut-scores. It is recommended that whenever possible, diverse perspectives be included in the composition of population samples, focus groups, subject matter expert groups, and expert judge panels. Clearly, the importance of diversity in such groups extends beyond sex, and may also include experience, ethnicity, and age.

Implications of nutrition and hydration

With the exception of hydration status, the nutritional status of the worker has been inadequately considered. Nutritional behaviours are important to long-term health, can impact the ability to work safely and effectively, and also impact on the performance of strenuous, prolonged physical and mental activity. Moreover, nutritional supplementation can be an important influence on performance, and needs to be considered within the context of evaluating individuals for their readiness for work. For example, the ingestion of caffeine increases mental alertness, reduces mental errors in many tasks, and improves physical performance and endurance (Shearer et al. 2016). To accurately evaluate a person's ability to perform a work-related task, the nutritional status of the person prior to testing should be evaluated and standardized. One must consider the possibility that supplementation in two otherwise healthy individuals of equivalent physiological attributes may enable one person to satisfy the cut-score and the other to be found deficient. This possibility needs to be considered and evaluated, both during the development of employment standards and in the derivation of the eventual cut-scores.

Implications of load carriage

Load carriage increases physiological strain during ambulatory activities and reduces work capacity by siphoning part of the energy reserve to support and move the load, leaving less energy for locomotion (Taylor et al. 2016). This situation can imply inequalities within occupations where workers must carry and work with the same equipment. The increase in the absolute oxygen cost of load carriage is generally proportional to the change in the overall load (body, clothing, equipment, and other carried masses), with smaller individuals consuming more energy per unit mass, regardless of the nature of that mass (Taylor et al. 1980). Consequently, smaller people require greater aerobic fitness to meet the acceptable employment standard whenever endurance-based aptitude tests are conducted using fixed absolute loads.

When characterizing the metabolic demands of work, previous investigators have attempted to reduce the possibility of a mass bias by normalizing data to yield a mass-specific quantity (e.g., relative oxygen consumption in $mL \cdot kg^{-1} \cdot min^{-1}$; Royal Society (Great Britain), Symbols Committee 1975). That procedure is often used when comparing people of different sizes (e.g., males versus females). Unfortunately, arithmetic normalizations, or ratio standards, are invalid unless the resulting regressions pass though the origin (Packard and Boardman 1999). An example of considerable relevance to employment standards occurs after normalizing oxygen consumption data, with physiologically impossible outcomes commonly observed; that is, a zero ordinate intercept (oxygen cost) is rarely evident when the body mass is zero (Tanner 1949; Taylor et al. 2016). It is therefore recommended that the metabolic cost of ambulatory workplace activities be reported in absolute units (McLellan and Havenith 2016; Taylor et al. 2015*c*, 2016).

The combined effect of the two previous points has led the current authors to challenge the validity of most unloaded, endurance tests used internationally for screening firefighters (e.g., unloaded graded exercise tests, shuttle-run tests, obstacle course tests). The use of unloaded tests, and the ensuing cut-scores for firefighters (maximal oxygen consumption at or near $45\ mL \cdot kg^{-1} \cdot min^{-1}$), was largely based on the classical observations of Gledhill and Jamnik (1992*b*) that were completed on loaded firefighters. Recent experimental observations and interpretations suggest a re-thinking of that convention. First, those observations did not necessarily provide a complete representation of the workplace demands (Taylor et al. 2015*c*) and, in some cases, the operational work intensity. Second, the recommended mass-specific cut-score did not incorporate the masses of the protective clothing, equipment, or carried loads (Taylor et al. 2016). Third, the possibility existed that the cut-score may have been influenced by the characteristics (body mass) of those tested (Taylor et al. 2016).

Importantly, while physiological maxima (e.g., peak heart rate, peak aerobic power (absolute: $L \cdot min^{-1}$)) are usually only slightly affected (2%–5% reduction) by load carriage (protective clothing and equipment, and heavy backpacks) up to approximately 25 kg, other indicators of exercise performance and work tolerance (e.g., exercise test duration, peak power output) are reduced more significantly (10%–50%: Louhevaara et al. 1995; Phillips et al. 2016; Taylor et al. 2012*a*). If the outcome of interest was simply the determination of peak aerobic power in absolute units, then either a loaded or unloaded test would provide much the same information. However, if the test results were to be used to determine work tolerance under load, then a loaded test is required. Therefore, it is recommended that loaded tests be used to assess readiness for work in occupations for which a significant load carriage is both critical and endurance-related.

The same reasoning applies to unloaded obstacle-course tests commonly used to evaluate fitness for police officers (Bonneau and Brown 1995; Bonneau 1996; Rhodes and Farenholtz 1992). Typically, those tests require completion of a series high-intensity activities with many changes in direction (e.g., running, jumping, vaulting, fight simulations) in a relatively short time (~4 min), with readiness for work then inferred from the test completion

time. The combination of intensity and duration indicates a reliance on both aerobic and anaerobic endurance. Front-line police officers are normally encumbered with equipment (e.g., duty belt, weapons, radio), body armour, uniform, and boots. Three groups have investigated the impact of various loads (8–23 kg) on the operational mobility of police officers (Carlton et al. 2014; Dempsey et al. 2013; Lewinski et al. 2015). In those studies, performance decrements ranged from 7%–34%, depending on the load and aspects of mobility evaluated. Those results suggest that the best method for evaluation of physiological readiness for work in police officers must involve loaded test protocols.

Accordingly, it is recommended that physiological aptitude tests for all jobs involving personal protective clothing and equipment loads should be performed while wearing those ensembles and carrying that equipment (McLellan and Havenith 2016; Taylor et al. 2016), or at least a reasonable approximation. Moreover, because the metabolic impact of load distribution varies by approximately 8- to 9-fold between loads borne on the feet and torso (Taylor et al. 2012*a*), it is further recommended that a faithful replication of the physiological impact of workplace loads must occur within those tests. For instance, adding a 2-kg mass to the torso, in an attempt to simulate footwear, is metabolically inappropriate. Equally unjustified is the adding of loads to match the mass, and to replace the wearing of protective clothing. While such practices replicate the absolute loads, they ignore the significant metabolic costs associated with locomotion and the location-specific metabolic burden of load carriage (Dorman and Havenith 2009; Taylor et al. 2012*a*). In summary, best practice dictates that all aspects of load carriage should be faithfully reproduced, to the extent that it is possible, during employment standards research and subsequent evaluations of the physiological readiness for work. In the event that such reproduction is not possible, then researchers are encouraged to undertake studies to, at the very least, identify the extent to which physiological responses differ between correctly loaded states and states of simulated load.

Thermal considerations

Occupational health and safety guidelines exist to protect workers from excessive, thermally mediated strain, since protracted strain can have an adverse influence on physiological performance. Thus, researchers must be aware of how best to quantify the physiological impact of diverse environments. One stress index was highlighted by Cheung et al. (2016), with regard to heat exposure: the wet-bulb globe temperature (WBGT) index. The authors drew attention to its limitations, as summarized by Budd (2008). Notwithstanding those qualifications and its ubiquitous workplace use, additional reservations concerning that index are provided below, but with a particular emphasis upon heat exposure within physically demanding occupations. This knowledge is relevant not just to the collection of data during the characterization of working duties (Table 1), but also to recruit screening, for the environmental impact can often represent one of the more physiologically demanding aspects of the job (Cheung et al. 2016; McLellan and Havenith 2016). With regard to the heat and cold-tolerance testing of potential recruits, readers are directed to the next section of this communication.

For readers unfamiliar with the different physical stress and physiological strain indices pertaining to the thermal environment, it is helpful to consider the following classifications. The stress indices were designed to quantify the thermal environment and its impact on thermal comfort. These are known as effective temperature (sensation) scales because they were derivations of the effective temperature scale developed for equating thermal comfort in air-conditioned offices (Houghten and Yagloglou 1923). Consequently, each derivation, including the WBGT index, comes with the limitations of that scale. Moreover, none of the sensation indices quantifies physiological strain (Wenzel et al. 1989). Indeed, while there is no doubt that conditions having a WBGT difference of 10 °C will impose different stress on the worker, there is no justification for suggesting that environments with the same WBGT will be equally stressful. Moreover, the further the working environment moves away from that of the air-conditioned office, the less valid will that index be for setting operating standards within the workplace. Thus, its utility is questionable, at best, for individuals wearing protective clothing, carrying loads, and performing hard physical work (Taylor 2006).

Alternatively, readers are directed to the rational scales, which are based on the heat balance equation. As such, those scales account for both the need for heat loss and the capacity of the environmental conditions to support the physical and physiological avenues for heat loss. One particularly useful scale is the heat strain index (HSI) developed by Belding and Hatch (1955). That index quantifies thermal compensability by incorporating factors in addition to the characteristics of the thermal environment; metabolic heat production, clothing insulation and moisture permeability, and required evaporation. Indeed, the HSI is preferred by most thermal physiologists. It is therefore recommended, when occupational physiologists seek to quantify the physiological impact of hot environments, that they do so using the HSI and not the WBGT index, or its various modifications.

In comparison with the heat, there are few indices for working in the cold. The most common is the wind-chill index, which reflects the cooling power of the environment (equivalent-chill temperature) and the danger to exposed flesh in terms of time to freeze. It is worth emphasizing that it is a physical impossibility for wind chill to lower air temperate below ambient temperature; that is, wind chill will lower tissue temperature closer to, but not below, ambient temperature. The temperatures quoted in a wind-chill table are those that would have to be encountered in still air to achieve the same cooling power as that experienced with different wind speeds at the current ambient temperature. Many workplaces lack maximum exposure limits for cold work, or base those limits on the wind-chill index or the work warm-up schedule, which is used to recommend work-to-rest (rewarming) schedules for different air temperatures and wind speeds. It is recommended that neither workplace evaluations nor recruitment testing be undertaken under excessively cold conditions, unless those conditions represent the true working environment.

In addition to peripheral freezing and nonfreezing cold injury, neuromuscular cooling can have a debilitating impact on worker performance, as well as increasing the likelihood of trips and falls. A current large area for worker compensation is nonfreezing cold injury, resulting from prolonged exposure to cold without tissue freezing, and it may have life-long and significant debilitating consequences. However, little is known about the pathogenesis and pathology of that condition. With profound neuromuscular cooling, physical activity becomes almost impossible. Despite this potentially large impact on worker safety and performance, little systematic work has been undertaken on the consequences of cold working environments on task performance, and the possible mitigating adjustments to employment standards to allow for such incapacitation.

Impact of personal protective clothing and equipment

Two more critical matters pertaining to protective clothing and equipment should be highlighted. The first relates to the suggestion that heat-tolerance testing be incorporated into relevant workplaces (McLellan and Havenith 2016); similarly, it may be suggested that cold-tolerance testing be considered for workers entering workplaces in which cold stress represents a significant and frequent physiological challenge. Since some workers wear heavy-duty, thermal protective clothing during hard physical work, it is not uncommon for such individuals to store significant amounts of heat, independently of the thermal environment (McLellan et al. 2013). When exposed to external heat sources, that problem is compounded, yet even if recruits were tested wearing

the full protective ensemble, typical physiological aptitude test durations would be insufficient to elicit thermally challenging heat storage. Therefore, to minimize the incidence of workplace heat illness, heat-tolerance testing has been proposed, possibly also for incorporation within some employment standards.

Examples of heat-tolerance tests are discussed elsewhere (Cheung et al. 2016), although analysis reveals the well-established association between aerobic fitness and thermal adaptation that is inherent within those tests (Taylor 2014), such that peak aerobic power is a powerful predictor of heat tolerance. Nevertheless, this suggestion requires further consideration: when could heat-tolerance testing most efficiently be undertaken? Some would propose that this might best be performed after the initial occupational screening, when the subset of potential workers has been trimmed to a more manageable size.

Second, in some workplaces, breathing apparatus is essential. These devices add mass to the overall protective ensemble, although this is negligible during underwater work, and they also add a ventilatory burden by modifying the elastic and flow-resistive work of breathing (Butcher et al. 2006, 2007; Nelson et al. 2009; Taylor and Morrison 1999). Several factors should be taken into consideration when evaluating this problem, including load carriage and exercise intensity. The added mass is relatively easy to evaluate, with modern breathing apparatus being relatively light and generally carried with well-designed backpacks. More complex is the potential flow-resistive breathing restriction, which is dependent on the ventilatory rate. Eves et al. (2005) and Butcher et al. (2006, 2007) showed that external breathing resistance from the secondary regulator leads to increased work of breathing and becomes significant when ventilation exceeds approximately 75 $L \cdot min^{-1}$. Thus, even during submaximal exercise, both the mass and breathing resistance add to the physiological strain. Eves et al. (2005) and Dreger et al. (2006) described the attenuation of minute ventilation at peak exercise when using breathing apparatus, which resulted in proportional reductions in peak oxygen uptake. During, submaximal and even moderate- to high-intensity work, the pulmonary reserve is likely to be sufficient to absorb the additional physiological strain without impairing work performance (Dreger and Petersen 2007). However, at maximal effort, physiological maxima and work performance must be reduced. McLellan and Havenith (2016) identified the possibility that individuals who marginally satisfy the cut-score of a physiological aptitude test performed when not using breathing apparatus, may prove to be deficient in the workplace when performing the same task using breathing apparatus. This important matter is again considered below.

Recommendations for future research

A recurring theme from the conference, and subsequently in the papers in this special issue, is the recognition that screening tests for employment are normally conducted under near-ideal conditions, while workers are frequently required to respond under far more severe conditions. Accounting for the potential differences between the conditions under which physiological readiness for work is assessed and the actual physical demands of the job must be a focus in employment standards research and related disciplines (e.g., environmental physiology, ergonomics, nutrition, occupational psychology). The implications of this concern were illustrated by Nindl et al. (2013) when discussing the physiological and medical consequences of international military deployments.

There is little doubt that age and sex are the most common grounds for discrimination in employment screening. Readers are referred to accompanying works by Kenny et al. (2016) and Roberts et al. (2016) for detailed treatments of these factors with regard to work capability. Despite their importance, surprisingly little emphasis in the employment standards, and the related disciplinary literature, has been placed on documenting actual differences and the effect of interventions on initial and retained work ability.

In the first case, much of the literature contains data obtained from convenience samples of males and females, or younger and older adults. Very little research has been undertaken between groups that were carefully matched for critical characteristics. As an example, Boyd et al. (2015) studied variability in performance on a test of fitness for duty for Canadian Forces firefighters in a convenience sample of male (n = 31) and female (n = 20) subjects. As expected, mean mass and stature were lower for the females and their mean test performances were slower. However, when male and female subgroups were matched on fitness and size, there were no sex-dependent differences in performance variability. More research is required that specifically addresses sex- and age-dependent differences through carefully designed experiments.

In the second case, more research is warranted on the effects of the specificity of occupationally relevant physical training interventions. As an example, Jamnik et al. (2010*c*) examined pass and fail rates on a test of readiness for work in potential applicants to the corrections service before and after a 6-week, test-specific training program. Not surprisingly, pass rates increased significantly, and this effect was substantially greater in the females. The authors found that outcome was largely dependent upon prior exercise behaviours, particularly in those with lower basal abilities within the attributes of interest. The benefit was most pronounced in females, but applies to both sexes. More research on physiological adaptations and task-related performance changes from similar interventions would make valuable contributions to the employment standards field of study, especially when such interventions can be used as accommodation strategies.

At present, there is only a broad understanding of the physiological consequences of load carriage; that is, most investigators have considered only the average (mostly male) members of the population. To expand our understanding, load-carriage research is required at both extremes of the adult size spectrum, with the following questions appearing to represent priority areas. Does the oxygen cost of locomotion vary as a power function of body mass, as it does in other species? Is that relationship altered when loads are added to the torso? Can those relationships explain the sex differences observed during these activities? Do the body-location oxygen costs of load carriage remain constant when evaluated in very small and very large individuals, or in men and women?

McLellan and Havenith (2016) recommended that research is required to address the impact of breathing apparatus on physiological aptitude test scores. Presumably, and based on previous research (Butcher et al. 2006, 2007; Dreger et al. 2006; Eves et al. 2005), scores for tests where aerobic fitness is a main limiting factor would decline. It is unknown how scores would be modified for tests in which strength and anaerobic fitness were the main limiting factors. If an affirmative outcome is realized, then it is necessary to know how employment standards and cut-scores should be modified to take the added burden and breathing resistance into consideration, if tests are to be performed without using breathing apparatus.

There is a growing body of evidence, some of which was reviewed by Cheung et al. (2016), pertaining to the thermal interaction of heat stress and cognition. This is an important area of research for occupational physiology. However, there is an accumulation of confounding evidence within this area, and there are three reasons why this has occurred. First, many experiments have been performed in relatively uncertain thermal states. Typically, ambient conditions were recorded, and presumably with precision. Unfortunately, steady-state, deep-body temperatures were not often established. Thus, an ability to interpret the resulting data with regard to changes in central nervous system temperature was frequently very limited. Second, because of the presence of very high performance scores within basal conditions, many experiments were designed with a predetermined outcome

bias; that is, only performance decrements could be observed. Third, when different cognitive domains were evaluated within the same experiment, tests were often administered at different levels of difficulty. Such a design renders the resulting outcomes almost uninterpretable. Clearly, this is a research area within which significant improvements can be made, and it is recommended that tighter inter-disciplinary collaboration may reap significant gains.

Individuals employed in public safety and protection environments are exposed to significant levels of nonphysical stress, and the response to each type of stress may vary greatly (Anderson et al. 2002). A common feature of situations that initiate a stress response is the association with some degree of unpredictability, either in terms of a situation or an outcome. The resulting physiological response triggers a complex hormonal cascade that arguably is one of the most potent influences on brain function and behaviour. There are few studies, however, that document the impact of stress on the performance of employment tests, and none that allow quantification of the impact of stress on a performance standard. What is known is that acute and chronic stress can change the execution of both fine and gross movements and balance (Metz et al. 2005). Muscle tension may also increase with high levels of stress (Lundberg et al. 1994), reducing movement efficiency (Mehta et al. 2012). These data support the expectation that decrements in skilled performance will occur under stress, even in simulated environments. There is also some evidence that aerobic fitness may modulate the autonomic nervous response to acute stress (von Haaren et al. 2016), thus providing a protective benefit. If this effect is found to hold following further investigation, it could provide evidence in support of elevated aerobic fitness requirements for workers in high-stress environments whose work does not normally involve a high aerobic demand (e.g., police officers). Nevertheless, the relationship between stress and employment standards has yet to be investigated.

Conclusions

There is a fundamental responsibility to ensure that performance standards, physiological aptitude tests, and cut-scores are able to identify those individuals who, on the balance of probabilities, can (or cannot) work safely and effectively. Incorrect decisions put employees at risk and jeopardize public safety. However, even with the most carefully crafted tests, standards, and cut-scores, there remains an element of uncertainty. It is recommended that researchers undertake steps to address those uncertainties. One approach would be to establish confidence intervals that reflect uncertainty within cut-scores. Prediction intervals provide greater certainty when used with the same data to set a population score (as opposed demonstrating sample variability) based on predictive data. Finally, it is important to recognize the inherent variability in human performance, and to evaluate its impact on test scores.

It is timely to align best practice in science with occupational practice to increase the probability of correct employee recruitment and retention decisions. Clarification of nomenclature is considered important. In particular, the terms standard and cut-score have often been used inconsistently. The authors have recommended definitions for these words that comply with the discipline of psychometrics, with the aim of bringing greater precision and clarity to this field.

Since performance declines with age, it is recommended that employees be tested on a more regular, and age-associated, basis to confirm their ability to work in a safe and effective manner. Similarly, tests of readiness for work should be considered, and where appropriate, used to assess employees returning to work after injury, prolonged sickness, and redeployment.

In the accompanying review papers, the need has been highlighted for a more multi-disciplinary approach to some of the critical research in this field. Environmental challenges, protective clothing, and load carriage are common in most physically demanding occupations. These factors add to the physiological strain already inherent in the working environment, and need to be considered when developing aptitude tests, standards, and cut-scores. Research is required to determine how the actual demands on the job can be factored into the testing environment, and how performance measured under less stressful conditions deteriorates under real-world conditions.

Although an under-researched and often ill-considered area, the development of valid and defensible employment standards has significant economic and social impact. Few areas of science carry the responsibility of allowing, or denying, an individual the opportunity to work. This responsibility underscores the priority that must be assigned to the support and funding of employment standards research to enable knowledge gaps to be filled and, as a consequence, valid and reliable ways to be developed for ensuring workers in physically demanding occupations are safe, healthy, fulfilled, and productive.

Conflict of interest statement

The authors of this manuscript report no conflicts of interest.

References

Adams, E.M. 2016. Human rights at work: physical employment standards and human rights law. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0552.

Altman, D.G., and Bland, J.M. 1983. Measurement in medicine: the analysis of method comparison studies. Statistician, **32**(3): 307–317. doi:10.2307/2987937.

Anderson, G.S., Plecas, D.B., and Litzenberger, R. 2002. Physical evidence of police officer stress. Pol. Int. J. Police Strategies Manage. **25**: 399–420. doi:10.1108/13639510210429437.

Belding, H.S., and Hatch, T.F. 1955. Index for evaluating heat stress in terms of resulting physiological strain. Heating, Piping and Air Conditioning. **27**: 129–136.

Blacker, S.D., Rayson, M.P., Wilkinson, D.M., Carter, J.M., Nevill, A.M., and Richmond, V.L. 2015. Physical employment standards for UK fire and rescue service personnel. Occup. Med. **66**(1): 38–45. PMID:26428442.

Blacklock, R.E., Reilly, T.J., Spivock, M., Newton, P.S., and Olinek, S.M. 2015. Standard establishment through scenarios (SETS): a new technique for occupational fitness standards. Work, **52**(2): 375–383. doi:10.3233/WOR-152128. PMID:26409372.

Blair, S.N., Horton, E., Leon, A.S., Lee, I.M., Drinkwater, B.L., Dishman, R.K., et al. 1996. Physical activity, nutrition, and chronic disease. Med. Sci. Sports Exerc. **28**(3): 335–349. doi:10.1249/00005768-199603000-00009. PMID:8776222.

Bland, J.M., and Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, **327**(8476): 307–310. doi:10.1016/S0140-6736(86)90837-8.

Bland, J.M., and Altman, D.G. 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res. **8**(2): 135–160. doi:10.1191/096228099673819272. PMID:10501650.

Bonneau, J. 1996. PARE Standard Summary. RCMP Health Services Directorate, Ottawa, Ont., Canada.

Bonneau, J., and Brown, J. 1995. Physical ability, fitness and police work. J. Clin. Forensic Med. **2**: 157–164. doi:10.1016/1353-1131(95)90085-3.

Booth, F.W., and Lees, S.J. 2006. Physically active subjects should be the control group. Med. Sci. Sports Exerc. **38**(3): 405–406. doi:10.1249/01.mss.0000205117.11882.65. PMID:16540824.

Booth, F.W., Gordon, S.E., Carlson, C.J., and Hamilton, M.T. 2000. Waging war on modern chronic diseases: primary prevention through exercise biology. J. Appl. Physiol. **88**(2): 774–787. PMID:10658050.

Booyens, J., and Hervey, G.R. 1960. The pulse rate as a means of measuring metabolic rate in man. Can. J. Biochem. Physiol. **38**(11): 1301–1309. doi:10.1139/o60-162.

Bowes, H.M., Burdon, C.A., and Taylor, N.A.S. 2015. The scaling of human basal metabolic rate in adult males. Proc. Aust. Physiol. Soc. **46**: 54P.

Boyd, L., Rogers, T., Docherty, D., and Petersen, S. 2015. Variability in performance on a work simulation test of physical fitness for firefighters. Appl. Physiol. Nutr. Metab. **40**(4): 364–370. doi:10.1139/apnm-2014-0281. PMID:25781347.

Budd, G.M. 2008. Wet-bulb globe temperature (WBGT) – its history and its limitations. J. Sci. Med. Sport, **11**(1): 20–32. PMID:17765661.

Butcher, S.J., Jones, R.L., Eves, N.D., and Petersen, S.R. 2006. Work of breathing is increased during exercise with the self-contained breathing apparatus regulator. Appl. Physiol. Nutr. Metab. **31**(6): 693–701. doi:10.1139/h06-073. PMID:17213883.

Butcher, S.J., Jones, R.L., Mayne, J.R., Hartley, T.C., and Petersen, S.R. 2007. Impaired exercise ventilatory mechanics with the self-contained breathing ap-

paratus are improved with heliox. Eur. J. Appl. Physiol. **101**(6): 659–669. doi:10.1007/s00421-007-0541-5. PMID:17701048.

Carlton, S.D., Carbone, P.D., Stierli, M., and Orr, R. 2014. The impact of occupational load carriage on the mobility of the tactical police officer. J. Aust. Strength Cond. **22**(1): 32–37.

Chakravarthy, M.V. 2008. Physiological penalties of the sedentary lifestyle. *In* Physiological Bases of Human Performance During Work and Exercise. *Edited by* N.A.S. Taylor and H. Groeller. Churchill Livingstone Elsevier, Edinburgh, UK. pp. 493–504.

Cheung, S.S., Lee, J.K.W., and Oksa, J. 2016. Thermal stress, human performance and physical employment standards. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0518.

Constable, S.H., and Palmer, B. 2000. The Process of Physical Fitness Standards Development. Human Systems Information Analysis Center Program Office, Wright Patterson Air Force Base, Ohio, USA.

Davis, P.O., Dotson, C.O., and Santa Maria, D.L. 1982. Relationship between simulated fire fighting tasks and physical performance measures. Med. Sci. Sports Exerc. **14**(1): 65–71. doi:10.1249/00005768-198201000-00013. PMID:7070261.

De Rey, C.C.M.-V., Galindo, M.P.G., and Velarde, M.A.A. 2001. Effects of using mean scores in regression models: an example from environmental psychology. Quality and Quantity, **35**(2): 191–202. doi:10.1023/A:1010384927696.

Dempsey, P.C., Handcock, P.J., and Rehrer, N.J. 2013. Impact of police body armour and equipment on mobility. Appl. Ergon. **44**(6): 957–961. doi:10.1016/j.apergo.2013.02.011. PMID:23668780.

Dorman, L.E., and Havenith, G. 2009. The effects of protective clothing on energy consumption during different activities. Eur. J. Appl. Physiol. **105**(3): 463–470. doi:10.1007/s00421-008-0924-2. PMID:19011890.

Dreger, R.W., and Petersen, S.R. 2007. Oxygen cost of the CF-DND fire fit test in males and females. Appl. Physiol. Nutr. Metab. **32**(3): 454–462. doi:10.1139/H07-020. PMID:17510680.

Dreger, R.W., Jones, R.L., and Petersen, S.R. 2006. Effects of the self-contained breathing apparatus and fire protective clothing on maximal oxygen uptake. Ergonomics, **49**(10): 911–920. doi:10.1080/00140130600667451. PMID:16803723.

Durnin, J.V.G.A., and Edwards, R.G. 1955. Pulmonary ventilation as an index of energy expenditure. Q. J. Exp. Physiol. Cogn. Med. Sci. **40**(4): 370–377. doi:10.1113/expphysiol.1955.sp001137.

Eid, E. 2001. Challenges posed by the Supreme Court of Canada in the Meiorin Decision to employers in physically demanding occupations. *In* Bona Fide Occupational Requirements. Proceedings of the Consensus Forum on Establishing *Bona Fide* Requirements for Physically Demanding Occupations, September 13th–16th, 2000. *Edited by* N. Gledhill, J. Bonneau, and A. Salmon. York University, Toronto, Ont., Canada. pp. 53–61.

Epstein, Y., Yanovich, R., Moran, D.S., and Heled, Y. 2013. Physiological employment standards IV: integration of women in combat units physiological and medical considerations. Eur. J. Appl. Physiol. **113**(11): 2673–2690. doi:10.1007/s00421-012-2558-7. PMID:23238928.

Eves, N.D., Jones, R.L., and Petersen, S.R. 2005. The influence of the self-contained breathing apparatus (SCBA) on ventilatory function and maximal exercise. Can. J. Appl. Physiol. **30**(5): 507–519. doi:10.1139/h05-137. PMID:16293900.

Friedl, K.E. 2016. Biases of the incumbents: What if we were integrating men into a women's Army? Mil. Rev. **96**(2): 69–75.

Fullagar, H.H.K., Sampson, J.A., Mott, B.J., Burdon, C.A., Taylor, N.A.S., and Groeller, H. 2015. Employment standards for Australian urban firefighters. Part 4: Physical aptitude tests and standards. J. Occup. Environ. Med. **57**(10): 1092–1097. doi:10.1097/JOM.0000000000000528. PMID:26461864.

Giavarina, D. 2015. Understanding Bland Altman analysis. Biochem. Med. **25**(2): 141–151. doi:10.11613/BM.2015.015.

Gledhill, N., and Bonneau, J. 2001. Objectives, process and consensus summary of the National Forum on *bona fide* occupational requirements. *In* Bona Fide Occupational Requirements. Proceedings of the Consensus Forum on Establishing *Bona Fide* Requirements for Physically Demanding Occupations, September 13th–16th, 2000. *Edited by* N. Gledhill, J. Bonneau, and A. Salmon. York University, Toronto, Ont., Canada. pp. 1–6.

Gledhill, N., and Jamnik, V.K. 1992*a*. Characterization of the physical demands of firefighting. Can. J. Sport Sci. **17**(3): 207–213. PMID:1325260.

Gledhill, N., and Jamnik, V.K. 1992*b*. Development and validation of a fitness screening protocol for firefighter applicants. Can. J. Sport Sci. **17**(3): 199–206. PMID:1325259.

Gledhill, N., and Jamnik, V. 2011. Report on the development and validation of a Canadian physical performance exchange standard for type 1 wildland forest fire fighters. Canadian Inter-agency Forest Fire Center Inc.

Goldman, R.F. 2001. Introduction to heat-related problems in military operations. *Edited by* K.B. Pandolf, R.E. Burr, C.B. Wenger, and R.S. Pozos. *In* Medical Aspects of Harsh Environments. Vol. 1. Textbook of military medicine. *Edited by* R. Zajtchuk and R.F. Bellamy. Department of the Army, Office of the Surgeon General, and Borden Institute. Washington, DC, USA. pp. 3–49.

Groeller, H. 2008. The physiology of ageing in active and sedentary humans. *In* Physiological Bases of Human Performance During Work and Exercise. *Edited by* N.A.S. Taylor and H. Groeller. Churchill Livingstone Elsevier, Edinburgh, UK. pp. 289–306.

Groeller, H., Fullagar, H.H.K., Sampson, J.A., Mott, B.J., and Taylor, N.A.S. 2015. Employment standards for Australian urban firefighters. Part 3: The transition from criterion task to test. J. Occup. Environ. Med. **57**(10): 1083–1091. doi:10.1097/JOM.0000000000000527. PMID:26461863.

Halligan, S., Altman, D.G., and Mallett, S. 2015. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur. Radiol. **25**(4): 932–939. doi:10.1007/s00330-014-3487-0. PMID:25599932.

Handoll, H.H.G., and Langhorne, P. 2015. In defence of reviews of small trials: underpinning the generation of evidence to inform practice. Cochrane Database of Systematic Reviews, **2015**(11). doi:10.1002/14651858.ED000106.

Hanley, J.A., and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, **143**(1): 29–36. doi:10.1148/radiology.143.1.7063747. PMID:7063747.

Hatfield, R. 2005. Duty to accommodate. Just Labour, **5**: 23–33.

Hogan, J., and Quigley, A.M. 1986. Physical standards for employment and the courts. Am. Psychol. **41**(11): 1193–1217. doi:10.1037/0003-066X.41.11.1193. PMID:3813183.

Holloszy, J.O., and Kohrt, W.M. 2011. Exercise. *In* Comprehensive Physiology. Supplement 28: Handbook of Physiology, Aging. 633–666. [First published in 1995.] doi:10.1002/cphy.cp110124.

Houghten, F.C., and Yagloglou, C.P. 1923. Determination of the comfort zone. J. Am. Soc. Heat. Ventil. Eng. **29**: 515–532.

Ilmarinen, J.E. 2001. Aging workers. Occup. Environ. Med. **58**(8): 546–552. doi:10.1136/oem.58.8.546. PMID:11452053.

Jackson, A. 1994. Preemployment physical evaluation. Exerc. Sport Sci. Rev. **22**: 53–90. PMID:7925553.

Jamnik, V.K., Thomas, S.G., Shaw, J.A., and Gledhill, N. 2010*a*. Identification and characterization of the critical physically demanding tasks encountered by correctional officers. Appl. Physiol. Nutr. Metab. **35**(1): 45–58. doi:10.1139/H09-121. PMID:20130666.

Jamnik, V.K., Thomas, S.G., Burr, J.F., and Gledhill, N. 2010*b*. Construction, validation, and derivation of performance standards for a fitness test for correctional officer applicants. Appl. Physiol. Nutr. Metab. **35**(1): 59–70. doi:10.1139/H09-122. PMID:20130667.

Jamnik, V.K., Thomas, S.G., and Gledhill, N. 2010*c*. Applying the Meiorin Decision requirements to the fitness test for correctional officer applicants; examining adverse impact and accommodation. Appl. Physiol. Nutr. Metab. **35**(1): 71–81. doi:10.1139/H09-123. PMID:20130668.

Jamnik, V., Gumienak, R., and Gledhill, N. 2013. Developing legally defensible physiological employment standards for prominent physically demanding public safety occupations: a Canadian perspective. Eur. J. Appl. Physiol. **113**(10): 2447–2457. doi:10.1007/s00421-013-2603-1. PMID:23494548.

Jenkins, A.B., and Plasqui, G. 2008. Exercise and disease states. *In* Physiological Bases of Human Performance During Work and Exercise. *Edited by* N.A.S. Taylor and H. Groeller. Churchill Livingstone Elsevier, Edinburgh, UK. pp. 521–533.

Kane, M. 1994. Validating the performance standards associated with passing scores. Rev. Edu. Res. **64**(3): 425–461. doi:10.3102/00346543064003425.

Kenny, G.P., Groeller, H., McGinn, R., and Flouris, A.D. 2016. Age, human performance and physical employment standards. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0483.

Knapik, J.J., Reynolds, K.L., and Harman, E. 2004. Soldier load carriage: historical, physiological, biomechanical, and medical aspects. Mil. Med. **169**(1): 45–56. doi:10.7205/MILMED.169.1.45. PMID:14964502.

Lewinski, W.J., Dysterheft, J.L., Dicks, N.D., and Pettitt, R.W. 2015. The influence of officer equipment and protection on short sprinting performance. Appl. Ergon. **47**: 65–71. doi:10.1016/j.apergo.2014.08.017. PMID:25479975.

Louhevaara, V., Ilmarinen, R., Griefahn, B., Künemund, C., and Mäkinen, H. 1995. Maximal physical work performance with European standard based fire-protective clothing system and equipment in relation to individual characteristics. Eur. J. Appl. Physiol. **71**(2–3): 223–229.

Lundberg, U., Kadefors, R., Melin, B., Palmerud, G., Hassmén, P., Engström, M., and Dohns, I.E. 1994. Psychophysiological stress and EMG activity of the trapezius muscle. Int. J. Behav. Med. **1**(4): 354–370. doi:10.1207/s15327558ijbm0104_5. PMID:16250795.

McLellan, T.M., and Havenith, G. 2016. Protective clothing ensembles and physical employment standards. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0474.

McLellan, T.M., Daanen, H.A.M., and Cheung, S.S. 2013. Encapsulated environment. Compr. Physiol. **3**(3): 1363–1391. PMID:23897690.

Mehta, R.K., Nussbaum, M.A., and Agnew, M.J. 2012. Muscle- and task-dependent responses to concurrent physical and mental workload during intermittent static work. Ergonomics, **55**(10): 1166–1179. doi:10.1080/00140139.2012.703695. PMID:22849301.

Metz, C.E. 1978. Basic principles of ROC analysis. Semin. Nucl. Med. **8**(4): 283–298. doi:10.1016/S0001-2998(78)80014-2.

Metz, G.A., Jadavji, N.M., and Smith, L.K. 2005. Modulation of motor function by stress: a novel concept of the effects of stress and corticosterone on behavior. Eur. J. Neurosci. **22**(5): 1190–1200. doi:10.1111/j.1460-9568.2005.04285.x. PMID:16176362.

Milligan, G.S., Zumbo, B.D., Reilly, T.J., and Tipton, M.J. 2016. Validity and reliability of physical employment standards. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0669.

Mulrow, C.D. 1994. Rationale for systematic reviews. Br. Med. J. **309**(6954): 597–599. doi:10.1136/bmj.309.6954.597.

Narici, M.V., Stewart, C.E., and di Prampero, P.E. 2008. Detraining, bed rest and adaptation to microgravity. *In* Physiological Bases of Human Performance During Work and Exercise. *Edited by* N.A.S. Taylor and H. Groeller. Churchill Livingstone Elsevier, Edinburgh, UK. pp. 149–166.

Nelson, M.D., Haykowsky, M.J., Mayne, J.R., Jones, R.L., and Petersen, S.R. 2009. Effects of self-contained breathing apparatus on ventricular function during strenuous exercise. J. Appl. Physiol. **106**(2): 395–402. PMID:19008481.

Nelson, M.D., Petersen, S.R., and Dlin, R.A. 2010. Effects of age and counseling on the cardiorespiratory response to graded exercise. Med. Sci. Sports Exerc. **42**(2): 255–264. PMID:19927033.

NFPA. 2014. Firefighter fatalities in the United States – 2014. nfpa.org/research/reports-and-statistics/the-fire-service/fatalities-and-injuries/firefighter-fatalities-in-the-united-states. [Accessed 27 November 2015.]

Nindl, B.C., Castellani, J.W., Warr, B.J., Sharp, M.A., Henning, P.C., Spiering, B.A., and Scofield, D.E. 2013. Physiological Employment Standards III: physiological challenges and consequences encountered during international military deployments. Eur. J. Appl. Physiol. **113**: 2655–2672. doi:10.1007/s00421-013-2591-1. PMID:23430237.

Notley, S.R., Fullagar, H.H.K., Lee, D.S., Matsuda-Nakamura, M., Peoples, G.E., and Taylor, N.A.S. 2014. Revisiting ventilatory and cardiovascular predictions of whole-body metabolic rate. J. Occup. Environ. Med. **56**(2): 214–223. doi:10.1097/JOM.0000000000000086. PMID:24451619.

Notley, S.R., Peoples, G.E., and Taylor, N.A.S. 2015. The utility of heart rate and minute ventilation as predictors of whole-body metabolic rate during occupational simulations involving load carriage. Ergonomics, **58**(10): 1671–1681. doi:10.1080/00140139.2015.1026406. PMID:25746518.

Notley, S.R., Park, J., Tagami, K., Ohnishi, N., and Taylor, N.A. 2016. Morphological dependency of cutaneous blood flow and sweating during compensable heat stress when heat-loss requirements are matched across participants. J. Appl. Physiol. (1985). [Epub ahead of print.] doi:10.1152/japplphysiol.00151.2016. PMID:27125845.

Packard, G.C., and Boardman, T.J. 1999. The use of percentages and size-specific indices to normalize physiological data for variation in body size: wasted time, wasted effort? Comp. Biochem. Physiol. Part A Mol. Integr. Physiol. **122**(1): 37–44. doi:10.1016/S1095-6433(98)10170-8.

Payne, W., and Harvey, J. 2010. A framework for the design and development of physical employment tests and standards. Ergonomics, **53**(7): 858–871. doi:10.1080/00140139.2010.489964. PMID:20582767.

Phillips, D.B., Stickland, M.K., Lesser, I.A., and Petersen, S.R. 2016. The effects of heavy load carriage on physiological responses to graded exercise. Eur. J. Appl. Physiol. **116**(2): 275–280. doi:10.1007/s00421-015-3280-z.

Pollock, R.D., Carter, S., Velloso, C.P., Duggal, N.A., Lord, J.M., Lazarus, N.R., and Harridge, S.D. 2015. An investigation into the relationship between age and physiological function in highly active older adults. J. Physiol. **593**(3): 657–680. doi:10.1113/jphysiol.2014.282863.

Poplin, G.S., Roe, D.J., Peate, W., Harris, R.B., and Burgess, J.L. 2014. The association of aerobic fitness with injuries in the fire service. Am. J. Epidemiol. **179**(2): 149–155. doi:10.1093/aje/kwt213. PMID:24186973.

Powell, K.E., and Blair, S.N. 1994. The public health burdens of sedentary living habits: theoretical but realistic estimates. Med. Sci. Sports Exerc. **26**(7): 851–856. PMID:7934758.

Proctor, D., and Joyner, M.J. 2008. Exercise and ageing: can the biological clock be stopped? *In* Physiological Bases of Human Performance During Work and Exercise. N.A.S. Taylor and H. Groeller. Churchill Livingstone Elsevier, Edinburgh, UK. pp. 313–319.

Proper, K.I., van den Heuvel, S.G., De Vroome, E.M., Hildebrandt, V.H., and Van der Beek, A.J. 2006. Dose-response relation between physical activity and sick leave. Br. J. Sports Med. **40**(2): 173–178. doi:10.1136/bjsm.2005.022327. PMID:16432007.

Rayson, M.P., Wilkinson, D.M., Carter, J.M., and Nevill, A.M. 2009. National Firefighter Selection Process. Development and Validation of National Firefighter Selection Tests: Physical Tests. Department for Communities and Local Government, UK.

Reilly, T., Wooler, A., and Tipton, M. 2006*a*. Occupational fitness standards for beach lifeguards. Phase 1: the physiological demands of beach lifeguarding. Occup. Med. **56**(1): 6–11. doi:10.1093/occmed/kqi169.

Reilly, T., Iggleden, C., Gennser, M., and Tipton, M. 2006*b*. Occupational fitness standards for beach lifeguards. Phase 2: the development of an easily administered fitness test. Occup. Med. **56**(1): 12–17. doi:10.1093/occmed/kqi168.

Rhodes, E.C., and Farenholtz, D.W. 1992. Police officer's physical abilities test compared to measures of physical fitness. Can. J. Sport Sci. **17**(3): 228–233. PMID:1325263.

Roberts, D., Gebhardt, D.L., Gaskill, S., Borges, A.S., Roy, T.C., and Sharp, M.A. 2016. Current considerations related to physiological differences between the sexes and physical employment standards. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0540.

Rogers, W.T., and Ricker, K.L. 2006. Establishing performance standards and setting cut-scores. Alberta J. Edu. Res. **52**(1): 16–24.

Rogers, W.T., Docherty, D., and Petersen, S. 2014. Establishment of performance standards and cut-score for the Canadian Forces Firefighter Physical Fitness Maintenance Evaluation (FF PFME). Ergonomics, **57**(11): 1750–1759. doi:10.1080/00140139.2014.943680. PMID:25102916.

Royal Society (Great Britain), Symbols Committee. 1975. Quantities, Units and Symbols: A Report. The Royal Society, London, UK.

Schmidt-Nielsen, K. 1984. Scaling: Why is animal size so important? Cambridge University Press, Cambridge, UK.

Schulzer, M. 1994. Diagnostic tests: a statistical review. Muscle Nerve, **17**(7): 815–819. doi:10.1002/mus.880170719. PMID:8008012.

Shearer, J., Graham, T.E., and Skinner, T.S. 2016. Nutra-Ergonomics: Influence of nutrition on physical employment standards and the health of workers. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0531.

Shephard, R.J. 1991. Occupational demand and human rights. Public safety officers and cardiorespiratory fitness. Sports Med. **12**(2): 94–109. PMID:1947536.

Shephard, R.J., and Bonneau, J. 2002. Assuring gender equity in recruitment standards for police officers. Can. J. Appl. Physiol. **27**(3): 263–295. doi:10.1139/h02-016.

Sothmann, M.S., Gebhardt, D.L., Baker, T.A., Kastello, G.M., and Sheppard, V.A. 2004. Performance requirements of physically strenuous occupations: validating minimum standards for muscular strength and endurance. Ergonomics, **47**(8): 864–875. doi:10.1080/00140130410001670372. PMID:15204279.

Stevenson, J.M., Bryant, J.T., Andrew, G.M., Smith, J.T., French, S.L., Thomson, J.M., and Deakin, J.M. 1992. Development of physical fitness standards for Canadian Armed Forces younger personnel. Can. J. Sport Sci. **17**(3): 214–221. PMID:1325261.

Storer, T.W., Dolezal, B.A., Abrazado, M.L., Smith, D.L., Batalin, M.A., Tseng, C.H., and Cooper, C.B. 2014. Firefighter health and fitness assessment: a call to action. J. Strength Cond. Res. **28**(3): 661–671. doi:10.1519/JSC.0b013e31829b54da. PMID:24566608.

Supreme Court of Canada; British Columbia (Public Service Employee Relations Commission v BCGSEU). 1999. 3 S. C. R 3 (Meiorin Decision).

Swets, J.A. 1986. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol. Bull. **99**(1): 100–117. doi:10.1037/0033-2909.99.1.100. PMID:3704032.

Tanner, J.M. 1949. Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. J. Appl. Physiol. **2**(1): 1–15. PMID:18133122.

Taylor, C.R., Heglund, N.C., McMahon, T.A., and Looney, T.R. 1980. Energetic cost of generating muscular force during running: a comparison of large and small animals. J. Exp. Biol. **86**(1): 9–18.

Taylor, N.A.S. 2006. Challenges to temperature regulation when working in hot environments. Indust. Health, **44**(3): 331–344. PMID:16922177.

Taylor, N.A.S. 2014. Human heat adaptation. Compr. Physiol. **4**(1): 325–365. PMID:24692142.

Taylor, N.A.S., and Billing, D.C. (*Editors*). 2012. Physiological and Physical Employment Standards I. Proceedings of the First Australian Conference on Physiological and Physical Employment Standards (November 27th–28th, 2012. Canberra, Australia). University Of Wollongong, Wollongong, Australia. pp. 1–109.

Taylor, N.A.S., and Groeller, H. 2003. Work-based physiological assessment of physically-demanding trades: a methodological overview. J. Physiol. Anthropol. **22**: 73–81. doi:10.2114/jpa.22.73. PMID:12672970.

Taylor, N.A.S., and Morrison, J.B. 1999. Static respiratory muscle work during immersion with positive and negative respiratory loading. J. Appl. Physiol. **87**(4): 1397–1403. PMID:10517770.

Taylor, N.A.S., and Patterson, M.J. 2016. Military clothing and protective material: protection at the limits of physiological regulation. *In* The Mechanobiology and Mechanophysiology of Military-Related Injuries: Studies in Mechanobiology, Tissue Engineering and Biomaterials. Vol. 19. *Edited by* A. Gefen and Y. Epstein. Springer–Verlag GmbH, Berlin, Germany. pp. 303–332. doi:10.1007/8415_2014_181.

Taylor, N.A.S., and Taylor, E.A. 2012. How much money could an Australian fire brigade save annually by increasing firefighter fitness? *In* Physiological and Physical Employment Standards I. Proceedings of the First Australian Conference on Physiological and Physical Employment Standards, Canberra, Australia. November 27th–28th, 2012. *Edited by* N.A.S. Taylor and D.C. Billing. University of Wollongong, Wollongong, Australia. pp. 82–83.

Taylor, N.A.S., Lewis, M.C., Notley, S.R., and Peoples, G.E. 2012*a*. A fractionation of the physiological burden of the personal protective equipment worn by firefighters. Eur. J. Appl. Physiol. **112**(8): 2913–2921. doi:10.1007/s00421-011-2267-7. PMID:22143844.

Taylor, N.A.S., van den Heuvel, A.M.J., Kerry, P., McGhee, S., Peoples, G.E., Brown, M.A., and Patterson, M.J. 2012*b*. Observations on saliva osmolality during progressive dehydration and partial rehydration. Eur. J. Appl. Physiol. **112**(9): 3227–3237. doi:10.1007/s00421-011-2299-z. PMID:22230919.

Taylor, N.A.S., Dodds, M.J., Taylor, E.A., and Donohoe, A.M. 2015*a*. A retrospective evaluation of injuries to Australian urban firefighters (2003 to 2012): injury types, locations and causal mechanisms. J. Occup. Environ. Med. **57**(7): 757–764. doi:10.1097/JOM.0000000000000438. PMID:26067214.

Taylor, N.A.S., Fullagar, H.H.K., Mott, B.J., Sampson, J.A., and Groeller, H. 2015*b*. Employment standards for Australian urban firefighters. Part 1: The essential, physically demanding tasks. J. Occup. Environ. Med. **57**(10): 1063–1071. doi:10.1097/JOM.0000000000000525. PMID:26461861.

Taylor, N.A.S., Fullagar, H.H.K., Sampson, J.A., Notley, S.R., Durley, S.D., Lee, D.S., and Groeller, H. 2015*c*. Employment standards for Australian urban firefighters. Part 2: The physiological demands and the criterion tasks. J. Occup. Environ. Med. **57**(10): 1072–1082. doi:10.1097/JOM.0000000000000526. PMID:26461862.

Taylor, N.A.S., Peoples, G.E., and Petersen, S.R. 2016. Load carriage, human performance, and employment standards. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0486.

Thorlund, K., Anema, A., and Mills, E. 2010. Interpreting meta-analysis according to the adequacy of sample size. An example using isoniazid chemoprophylaxis for tuberculosis in purified protein derivative negative HIV-infected individuals. Clin. Epidemiol. **2**: 57–66. PMID:20865104.

Tipton, M.J., Milligan, G.S., and Reilly, T.J. 2013. Physiological employment standards I. Occupational fitness standards: objectively subjective? Eur. J. Appl. Physiol. **113**(10): 2435–2446. doi:10.1007/s00421-012-2569-4. PMID:23263741.

Turner, R.M., Bird, S.M., and Higgins, J.P.T. 2013. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane Reviews. PLoS ONE, **8**(3): e59202. doi:10.1371/journal.pone.0059202. PMID:23544056.

US Equal Employment Opportunity Commission. 1967. The Age Discrimination in Employment Act of 1967. eeoc.gov/laws/statutes/adea.cfm. [Accessed 8 November 2015.]

van der Ploeg, H.P., and Bauman, A.R. 2008. Thirty minutes of exercise: is it sufficient for health? *In* Physiological Bases of Human Performance During Work and Exercise. *Edited by* N.A.S. Taylor and H. Groeller. Churchill Livingstone Elsevier, Edinburgh, UK. pp. 309–312.

Vanagas, G. 2004. Receiver operating characteristic curves and comparison of cardiac surgery risk stratification systems. Int. Cardiovasc. Thoracic Surg. **3**(2): 319–322. doi:10.1016/j.icvts.2004.01.008.

von Haaren, B., Ottenbacher, J., Muenz, J., Neumann, R., Boes, K., and Ebner-Priemer, U. 2016. Does a 20-week aerobic exercise training programme increase our capabilities to buffer real-life stressors? A randomized, controlled trial using ambulatory assessment. Eur. J. Appl. Physiol. **116**: 383–394. PMID:26582310.

Walker, A., Driller, M., Argus, C., Cooke, J., and Rattray, B. 2014. The ageing Australian firefighter: an argument for age-based recruitment and fitness standards for urban fire services. Ergonomics, **57**(4): 612–621. doi:10.1080/00140139.2014.887790. PMID:24588283.

Wenzel, H.G., Mehnert, C., and Schwarzenau, P. 1989. Evaluation of tolerance limits for humans under heat stress and the problems involved. Scand. J. Work Environ. Health, **15**(Suppl. 1): 7–14. PMID:2609123.

Zumbo, B.D. 2016. Standard-setting methodology: establishing performance standards and setting cut scores to assist score interpretation. Appl. Physiol. Nutr. Metab. **41**: This issue. doi:10.1139/apnm-2015-0522.