NRC Research Press

REVIEW

Validity and reliability of physical employment standards[1]

Gemma S. Milligan, Tara J. Reilly, Bruno D. Zumbo, and Michael J. Tipton

Abstract: In this paper the role of validity and reliability in the development of physical employment standards (PESs) and the consideration of these factors in determining the final pass/fail criteria for a PES and ultimately the legal defensibility of a PES is examined. Particular attention is paid to the use of subject-matter experts, the levels of evidence used in the establishment of the minimum acceptable pace/intensity for the completion of critical tasks, and the considerations needed in physical test selection.

Key words: cut-scores, readiness for work, subject matter experts, minimum performance standard, levels of evidence.

Résumé : Dans cet article, nous analysons le rôle de la validité et de la fiabilité dans l'élaboration des normes physiques relatives à l'emploi (« PES »), la prise en compte de ces facteurs dans la prise de décision finale de réussite/échec d'une PES et, finalement, la solidité juridique d'une PES. Nous portons une attention particulière à l'utilisation des experts en la matière, aux niveaux de preuve utilisée dans la détermination de l'intensité minimale/rythme minimal acceptable pour la réalisation totale des tâches critiques et aux facteurs exigés dans la sélection du test physique. [Traduit par la Rédaction]

Mots-clés : seuils de coupure, aptitude au travail, experts en la matière, norme minimale de performance, niveaux de preuve.

Introduction

A physical employment standard (PES) may be challenged within a court of law; an expert witness can be called to determine the validity and reliability of a PES and whether it is rationally connected to the performance of the job (Tipton et al. 2013). Thus, validity and reliability are essential interrelated components of a PES, and an important consideration for those wishing to implement defensible cut-scores.

The development of a PES has previously been described as involving 6 methodological steps (Tipton et al. 2013); during each step there is a need to ensure that the process has been validated to progress onto the next step. To ensure clarity in the subsequent sections, Table 1 defines the common terminology used in the development of a PES.

Contemporary viewpoints on validity have emphasized "sources of validity evidence" rather than "types of validity", thereby focusing on the evidence for validity rather than simply its classification. For example, the *Test Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA)1999, 2014) applied within the educational system describe 5 sources of validity evidence based on (*i*) test content, (*ii*) response processes, (*iii*) internal structure, (*iv*) relations to other variables, and (*v*) the consequences of testing. Although it is a concept not widely used in the PES research literature, where possible, the terminology presented in the *Test Standards* will be used. Likewise, contemporary validity theory takes an integrative view of validity and reliability evidence. There has been some debate in the statistical psychometric literature as to whether reliability is a necessary but not sufficient condition for validity (Zumbo 2007). This issue is better cast as one of measurement precision so that one strives to have as little error as possible in measurement and inference. Specifically, reliability is a question of data quality, whereas validity is a question of inferential quality (Zumbo and Rupp 2004). As such, reliability and validity theory are interconnected research areas, and quantities derived in the former bound or limit the inferences in the latter, i.e., reliability is integral to validity in that a selection test or PES cannot be considered valid if it is not reliable.

The primary focus of this paper is validity and reliability in the context of the development of a legally defensible PES. Validity evidence will be addressed from 4 perspectives: content, logical, criterion, and construct, whilst reliability will be discussed from the perspectives of systematic and random error.

Evidence based on content (content validity) is the "degree to which a test adequately samples what was covered by the course" (Thomas and Nelson 2001). In the case of a PES, the "course" represents the critical tasks required by the job. Content validity cannot be determined through parametric quantitative analysis; however, nonparametric qualitative measures such as lists of specifications and requirements can be produced (Sireci 1998). The results of the critical task analysis and the method of best practice (MOBP) of these tasks should be related to the final PES.

Received 3 December 2015. Accepted 31 March 2016.

G.S. Milligan and M.J. Tipton. Extreme Environments Laboratory, Department of Sport and Exercise Sciences, University of Portsmouth, Spinnaker Building, Cambridge Road, Portsmouth, Hants PO1 2ER, UK.
T.J. Reilly. CFWMS Human Performance Research and Development Canadian Armed Forces, 4210 Labelle St., Ottawa, ON K1A 0K2, Canada.
B.D. Zumbo. Department of Educational and Counselling Psychology, and Special Education, University of British Columbia, Scarfe Building, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada.

Corresponding author: Gemma S. Milligan (email: gemma.milligan@port.ac.uk).

[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.

Table 1. Definitions for the common terminology used in the development of a physical employment standard.

	Abbreviation	Description
Validity	na	Degree to which the test measures what it is supposed to measure (Thomas and Nelson 2001)
Reliability	na	Constancy of a test to yield the same results (Thomas and Nelson 2001)
Critical task	na	Most critical and physically demanding "essential" components of the job (Tipton et al. 2013)
Subject matter expert	SME	An incumbent and or supervisor with experience and thorough knowledge of a task (Blacklock et al. 2015)
Method of best practice	MOBP	Standardized method by which a task should be performed, and is quantified in terms of task duration, rate, load, and technique (Tipton et al. 2013)
Minimum performance standard	MPS	Minimally adequate level of performance to perform the critical task
Direct task simulation	DTS	A valid simulation of the critical task
Predictive selection test	PST	A simple-to-measure test that adequately predicts performance of the critical tasks
Cut-score	na	The "passing score" of a PST or DST

Note: na, not applicable.

Logical validity, more commonly known as "face" validity, is achieved when a task analysis includes consultation with subject-matter experts (SMEs), experienced supervisors, and employees, and is most apparent when direct task simulations (DTS) are used as selection tests in developing PESs. Logical validity is strongly favoured by the courts when determining a PES, it is perceived that DTS are "job-related" (Berkman v. City of New York 1983; Henderson et al. 2007), whereas a predictive selection test (PST) could appear unrelated to the job and almost inevitably includes inherent error, relying on predictive relationships (Henderson et al. 2007; Tipton et al. 2013). Logical validity does not require the test developers to carry out criterion-based validation studies (Henderson et al. 2007). As evidence is based on subjective assessment of content (e.g., questionnaires), non-objective statistical evidence can be provided for logical validity thus, whilst logical validity may be considered important by an organization that wishes to promote the use of a PES, researchers tend to prefer a more objective measure of validity (Thomas and Nelson 2001), such as criterion validity.

There are 2 types of criterion-related evidence: concurrent and predictive. Concurrent validity is usually employed when a test is to be substituted by a simple or easily administered alternative; an example of this is the use of indirect assessments of maximum oxygen uptake (e.g., shuttle runs or step tests) as a valid replacement for the direct laboratory assessment (Siconolfi et al. 1985; Chatterjee et al. 2004; Sykes and Roberts 2004; McArdle et al. 2007). The rationale for the use of indirect tests include ease of administration, expense, more robust, and achievable within the constraints of the workplace, which often excludes the use of tests requiring expensive equipment. To determine whether a test has concurrent validity the correlation between the 2 methods is assessed: if a relationship exists (0.36 to 0.67 moderate; 0.68 and above strong (Taylor 1990)), the simpler test can be used (Reilly et al. 1979; Thomas and Nelson 2001). Whether or not such a relationship forms a defensible basis for determining employment should depend on additional considerations (discussed later).

The second type of criterion validity is predictive validity; this is especially important for determining the predictive capability of a test but this validity check is often neglected or not considered part of the research brief. In the case of a PES, this translates into the accuracy with which a selection test determines the capability of a previously untested individual to undertake a critical task required by the job. A correlation is specific to the population on which it was based, and applied to a different sample (potential workforce/applicants) may not be as accurate, thus lowering the validity coefficient, a tendency known as "shrinkage" (Thomas et al. 2005). One way to determine shrinkage is to assess the relevance of a prediction equation for a new sample drawn from the same population, that is, "cross-validation". It should be noted that shrinkage and cross-validation are not exclusive to criterion validity and should be considered for each of the validity methods.

Construct validity determines whether a PST measures the same constructs as those that actually govern the physical performance of the critical task. It is made up of convergent validity (i.e., constructs that theoretically should be related to each other are, in fact, observed to be related to each other) and discriminant validity (i.e., constructs that theoretically should not be related to each other are, in fact, observed to not be related to each other). Neither convergent or discriminant validity on their own are sufficient for establishing construct validity; both must be evident for a test to be valid. Thus, construct validity uses correlation to determine relationships/associations between constructs, e.g., someone with high–grip-strength endurance also performs well on a task such as load carriage (Reilly 2007). All the forms of validity discussed above are used as evidence to support construct-related validity: a PES demonstrates construct validity if it differentiates between those individuals who are, and are not, capable of performing a critical task to the minimum performance standard (MPS).

Reliability refers from a measure of consistency (reproducibility) within the data; for example in a PES it could be related to the equipment used to determine the physical demands of a critical task or the tests that make up a PES. It is usual to determine reliability using the test–retest method, in which the first measure is compared with a second or third conducted using the same participants and conditions (Vincent and Weir 2005). In addition to test–retest, several additional factors can affect test reliability, these include consistency:

1. between testers (inter-rater), e.g., different SMEs provide similar opinions regarding critical task selection and minimum performance standards;
2. of participants (intra-subject), e.g., participants have been familiarized (systematic variability) with the methods of best practice needed to perform the critical task and are in the same physical state undertaking test–retests (biological variability) (Boyd et al. 2015);
3. of tester's performance (intra-rater), e.g., an SME asked to determine the minimum acceptable pace to perform a critical task will provide the same answer each time they are questioned.

Ensuring validity within a PES

The first stage in developing a legally defensible PES is a task analysis to identify job-related critical tasks and determine the MPS (Rayson 2000; Reilly et al. 2006*a*; Reilly 2007; Jamnik et al. 2010; Milligan 2013; Tipton et al. 2013; Siddall et al. 2014; Taylor et al. 2015*a*). It is well established that a task analysis provides the foundation to developing a legally defensible PES (Rayson 2000;

Table 2. Criteria for identifying task subject matter experts for military tasks.

1. Experience performing the task during military **exercise or training**
2. Experience performing the task during military **deployment domestically**
3. Experience performing the task during military **deployment internationally**
4. Experience performing the task during an **emergency situation**
5. Experience in a position of leadership where you have directed subordinates to perform the task and have observed the task being performed
6. Have witnessed the task being performed in an acceptable manner
7. Have witnessed the task being performed unsuccessfully and can attest to the reasons for, and the consequences of, this failure (*e.g.*: Person was not fit enough to drag a casualty to cover, therefore requiring that another soldier cease providing covering fire and assist)
8. Experience witnessing and/or performing the task using several techniques and can comment on the advantages and disadvantages of these techniques
9. Experience delivering formal training on the task (*e.g.*: teaching courses, developing training curricula, *etc.*)

Note: From Blacklock et al. 2015, reproduced with permission of Work (Reading, Mass.), Vol. 52, p. 377, © 2015 IOS Press.

Gledhill et al. 2001; Taylor and Groeller 2003; Jamnik et al. 2010). It is critical that an objective and scientific task analysis is performed to ensure content validity of the tasks chosen (Blacklock et al. 2015; Taylor et al. 2015*a*).

Use of SMEs

A task analysis can employ both subjective and objective methodologies (Tipton et al. 2013). SMEs have been used as a method of determining critical tasks (Reilly 2007; Jamnik et al. 2010; Milligan 2013; Rogers et al. 2014; Siddall et al. 2014; Blacklock et al. 2015; Taylor et al. 2015*a*). An SME has been widely accepted as an incumbent and or supervisor with experience and thorough knowledge of a task. Further subclassifications have been suggested to include expert judgements SMEs (e.g., research scientists or policy specialists), experiential experts (e.g., firefighters and other relevant practitioners), or representatives of key groups (e.g., female police officers or aboriginal community members who are also forest firefighters). For the purposes of this paper, the widely accepted definition detailed above of an SME will be used.

An SME panel should comprise a range of experts who have relevant and current job-specific experience and are able to provide detailed technical knowledge of job requirements. These experts should come from different areas of the job and represent a wide range of expertise (Siddall et al. 2014; Blacklock et al. 2015; Taylor et al. 2015*a*). The Canadian Armed Forces have set criteria in the selection of SMEs, whereby an SME is identified by their ability to meet at least 2 of the 9 criteria outlined in Table 2 (Blacklock et al. 2015).

The experience associated with the task determines the quality, and therefore validity, of an SME. As a group, all 9 criteria must be represented within an SME panel, satisfying all the necessary areas of expertise (Blacklock et al. 2015). It is recommended that SMEs represent task supervisors and incumbents across a range of ages, seniorities (e.g., rank), occupations, and sex (Lavin et al. 2007; Rayson 2000; Bonneau 2001).

SMEs are often employed to not only determine the critical tasks, but also to describe and confirm the MOBP for their execution, including the minimum acceptable pace/intensity of a task (Bilzon et al. 2002; Reilly and Tipton 2005; Phillips et al. 2012; Milligan 2013; Siddall et al. 2014; Rogers et al. 2014; Blacklock et al. 2015; Taylor et al. 2015*a*).

Selection of an SME panel is critical as some participants may draw from personal experience with the job (specifically their critical tasks) and, depending on their experience or expectations, they may describe tasks and their related demands with unintentional bias. Perhaps an SME believes that the PES of today should be as difficult as it was when they enlisted, not recognizing that the demands of the job may have been very different in the past due, for example, to improved technology. Therefore, when selected appropriately, SMEs provide a PES with logical validity ensuring, in part, that the Meiorin Test objectives are fulfilled (British Columbia Public Service Employee Relations Commission v BCGSEU (1999) 3 S.C.R.3). For this reason, failure to select appropriate SMEs could invalidate a task analysis and the resulting PES.

In some instances the MOBP cannot be established by the SME because of variations in practices within an organization. If this is the case, methods should be established to determine the most efficient way to undertake a task to ensure that subsequent DTS have content validity (Milligan 2013). These include the physical and physiological measurement of the different methods and equipment used to undertake the critical task.

Setting an MPS

Having established the critical tasks, the next crucial step in a task analysis is the establishment of the minimum acceptable pace/intensity for the completion of each critical task. This requirement ultimately determines the final pass/fail criteria for a PES (Tipton et al. 2013; Zumbo 2016). Therefore, how MPSs are identified and justified provides the foundation upon which all subsequent assessments of validity are made, and thereby determines the defensibility of a PES.

It is important to distinguish between "standard-setting" and determining a cut-score for a particular test. Kane (1994) distinguished the cut-score (which he refers to as a "passing score"), as a point on the test score scale, from the performance standard, which is the minimally adequate level of performance for some purpose (Zumbo 2016). This section will focus on the validity of setting MPS.

The methods used to determine MPSs are likely to be a focus of any challenges to a PES (Berkman v. City of New York 1983). Those setting a PES based on minimum performance should be able to justify "the technique required and the slowest rate regarded as acceptable in terms of health, safety, capability and work performance" (Tipton et al. 2013). Basing a PES on the minimum acceptable performance to undertake a critical task should mean it is independent of sex and age (Reilly and Tipton 2005; Epstein et al. 2013; Tipton et al. 2013), making it more defensible (Reilly and Tipton 2005; Milligan 2013; Tipton et al. 2013).

The evidence for establishing the MPS of a PES tends to be acquired through quasi-objective or subjective (SME) methods, or a combination of the two. It is illuminating to compare the level of evidence used to establish the MPS of a PES with that widely used by the scientific and medical community (e.g., the Scottish Intercollegiate Guidelines Network criteria for the grading of literature and procedures (sign.ac.uk/index.html); levels of evidence (Eccles and Mason 2001)); these criteria are presented in Table 3. These criteria place a large emphasis on systematic, randomized controlled trials with very low risk of bias; evidence produced by objective, quantitative means are ranked higher than other forms of evidence that rely on subjective, qualitative data collection. Given the potential importance of the consequences of performance on a PES, it might be argued that a similar approach is appropriate for the development of a PES.

Table 4 presents some examples of MPSs judged against the level of evidence presented in Table 3. This demonstrates that the highest level of evidence achieved by any PES is 2+/III (Tables 3

Table 3. Levels of evidence.

SIGN criteria		Eccles and Mason (2001)	
Level	Description	Level	Description
1++	High-quality meta-analyses, systematic reviews of RCTs, or RCTs with very low risk of bias	Ia	Evidence from meta-analysis of randomized controlled trials
1+	Well-conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias	Ib	Evidence from at least 1 randomized controlled trial
1–	Meta-analyses, systematic reviews, or RCTs with a high risk of bias	IIa	Evidence from at least 1 controlled study without randomization
2++	High-quality systematic reviews of case-control or cohort or studies High-quality case-control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal	IIb	Evidence from at least 1 other type of quasi-experimental study
2+	Well-conducted case-control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal	III	Evidence from nonexperimental descriptive studies, such as comparative studies, correlation studies, and case-control studies
2–	Case-control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal	IV	Evidence from expert committee reports or opinions and/or clinical experience of respected authorities
3	Nonanalytic studies; e.g., case reports, case series		
4	Expert opinion		

Note: Network criteria for the grading of literature and procedures available from sign.ac.uk/index.html, levels of evidence available from Eccles and Mason (2001). RCT, randomized control trial.

Table 4. Justifications used in setting existing minimum performance standards.

SIGN level	Eccles and Mason (2001) level	Level of objectivity	PES example
1++	Ia	1	No examples available
1+	Ib	1	No examples available
1–	IIa	1	No examples available
2++	IIb	1	No examples available
2+	III	1	*PES* – RNLI Lifeboat Crew (Reilly 2007) *Critical tasks and minimum pace/intensity/load* – Casualty recovery, recover a 70-kg casualty from the water as a 2-man lift *How the minimum pace/intensity/load was established* – 70-kg was the average bodyweight of UK males and females aged between 19 and 65 y (Pheasant and Haslegrave 2005). Individual requirement of a 35-kg lift in order to make an equal contribution to the overall task *Notes* – Elements of this example can be categorized as levels 4 and IV
2–	IV		No examples available
3		2	*PES* – Canadian Forces Firefighters (Rogers et al. 2014) *Critical tasks and minimum pace/intensity/load* – 1-arm hose carry, 1 hand, carry 15.24-m section of rolled 65-mm rubber jacketed hose (16.5 kg) 15.24 m, return the same distance carrying the hose in the other hand; ladder carry and raise, carry a 3.6-m roof ladder (13.6 kg) 15.24 m and raises it to a secure position against a wall. Charged hose drag, dragging 2 charged lengths of 44-mm hose a distance of 30.48 m. Ladder climb 1, using a 7.2-m ladder, 10-rung climb (3.45 m) up and down, 3 times. Weighted sled pull, pull a 16-mm static nylon rope attached to a weighted sled 15.24 m using a hand–overhand movement; walk 15.24 m to the starting position of the sled repeat the pull. Forcible entry, using a 4.5 kg steel-head sledge hammer, hit a target on a mechanical apparatus until a buzzer sounds. Victim rescue, walking backward, drag an 80-kg mannequin a distance of 26 m. Ladder climb 2, using a 7.2-m ladder, 10-rung climb (3.45 m) up and down, 2 times. Ladder lower and carry, lower and carry a 3.6-m aluminum roof ladder (13.6 kg) 15.24 m. Equipment carry, carry a tricep bar with weight plates and collars (total weight 36.4 kg) 15.24 m and return *How the minimum pace/intensity/load was established* – Minimum pace set by SMEs using video analysis, blinded voting (see main body of text for further explanation) *Notes* – Both these examples have elements which can be categorized as levels 4 and IV
4		3	*PES* – Oil and gas industry (Milligan 2013) *Critical tasks and minimum pace/intensity/load* – Stair climbing, climb a flight of stairs at a rate of 80 steps·min^{-1}. Ladder climbing, ERT ladder climb at 34.5 rungs·min^{-1} *How the minimum pace/intensity/load was established* – Established by SMEs

Note: Four levels have been assigned to the examples (1, objective; 2, objective + subjective; 3, subjective). The table also presents the representative Scottish Intercollegiate Guidelines Network (SIGN) and levels of evidence criteria reported in Table 3. ERT, emergency response team; PES, physical employment standard; RNLI, Royal National Lifeboat Institution; SME, subject-matter expert.

and 4). From a review of the literature, the most commonly reported level of evidence used in the development of an MPS was 4/IV (Table 3). For example, Reilly et al. (2006*a*) determined the minimum pace that a beach lifeguard should be able to paddle 300 m, or swim 200 m, on the basis of the time it could take for a casualty to drown and the likelihood of a successful resuscitation based on published case studies (Table 3). This utilized the work of Fainer et al. (1951), Conn et al. (1995), Golden and Tipton (2002), and others, which suggested that it takes approximately 2 min for a casualty to drown having been face down in water, and resuscitation is unlikely after 10 min of anoxia (Quan et al. 2014). Reilly et al. (2006*a*) concluded that the 2-min window is extended if the casualty is initially observed experiencing difficulties, thereby increasing the potential rescue time to a 3- to 4-min rescue window, with no more than 10 min before the casualty should be in a position to receive resuscitation. These criteria determined the swim, paddle, and tow performance requirements for United Kingdom beach lifeguards. With groups like the emergency services, the levels of evidence available rank higher (Tables 3 and 4) and the rationale for the determination of MPSs is more straightforward because of the available evidence on factors such as how quickly a fire spreads, how quickly someone heats up or cools down, and post-trauma survival times (Table 4). However, even in these cases the rationale, logic, and evidence on which PESs are based are not beyond challenge.

It becomes more difficult in situations with less objective demands; for example, where one is trying to determine the slowest pace it is acceptable for a person to climb a ladder or complete a mission (Table 4). In such cases MPSs tend to be based solely on the subjective opinion of SMEs (Table 4) (Rayson 1998; Bilzon et al. 2002; Reilly 2007; Milligan 2013; Siddall et al. 2014; Taylor et al. 2015*a*). Siddall et al. (2014) and Rogers et al. (2014) have suggested similar processes using an adapted "Bookmark method" to set MPSs in an attempt to reduce the subjective nature of SMEs (Table 4). Siddall et al. (2014) asked experienced training instructors to complete the critical tasks at their own pace. The average speed was then calculated and used as a central reference for deciding a "slow" and "fast" speed for each task. Subsequent completion of the critical tasks at the 3 speeds (slow, average, fast) were filmed and shown to the SME along with a contextualization of the critical task; i.e., what is involved in the critical task and the MOBP. SMEs voted anonymously for which speed was, in their opinion, the minimum acceptable pace for the simulated critical task; SMEs were also given the chance to choose a pace for the task that lay halfway between those displayed in each of the 3 videos, giving a choice of 5 rates/intensities. Results were then discussed by the SME and a group consensus arrived at. Similarly, the methodology recommended by Rogers et al. (2014) asked 25 SMEs to rate the speed of critical tasks as acceptable, unacceptable, and minimally acceptable; each of these speeds were then sped up and slowed down by 20 s to give 9 options. The SMEs were asked to bookmark between the simulations considered acceptable and unacceptable, which was followed by 3 further stages to build consensus, to set a suitable cut off score. Whilst these procedures are still largely subjective and rely on norm referencing, they do at least attempt to standardize the method used by SMEs to determine a minimum pace and provide a more structured methodology to obtaining logical and content validity.

It has been reported that none of the levels of evidence detailed in Table 3 are appropriate for all settings (Morley et al. 2010). If an MPS cannot be clearly identified and justified on a rational basis, it is doubtful whether the associated task should be included in a PES. It is recommended that researchers and organizations wishing to develop MPSs give serious consideration to the level of evidence used in their development to ensure the validity and defensibility of the resultant PES. There will be some tasks for which no MPS can be established; these tasks cannot be assessed by a PES. Thus, whilst the level of evidence approach is not necessarily applicable to the determination of MPSs, such an analysis does help contextualize and highlight the level of evidence supporting a PES. The extent to which SMEs are the "default methodology" for establishing MPSs and are often used in preference to seeking more objective criteria is an area requiring further consideration. This is particularly the case given that common use of SMEs result in most PESs being based on evidence that in some other disciplines is regarded as "weak".

The role of the employing organization

Ultimately an MPS must be agreed to and adopted by the employing organization. There is little in the literature to help determine who in the organization should provide this agreement. In Canada this has often taken the form of a project management team, including members or representatives from policy, legal, equity, human resources and human rights, training development, unions, and management (Gledhill et al, 2001). In producing this review, 2 United Kindgom organizations (the Royal National Lifeboat Institution and the Maritime and Coastguard Agency), with current PESs, were consulted to discuss who was tasked with accepting MPSs. The general consensus was that such standards would be approved for adoption via a board of governance or technical committee, thereby adhering to the Meiorin Test criteria that states firstly: "The employer must show that it adopted the standard for a purpose rationally connected to the performance of the job." Second, the employer must establish that it adopted the particular MPS in an honest and good faith belief that it is necessary for the fulfilment of that legitimate work-related purpose. Third, the employer must establish that the MPS is reasonably necessary to the accomplishment of that legitimate work-related purpose. To show that the standard is reasonably necessary, it must be demonstrated that it is impossible to accommodate individual employees sharing the characteristics of the claimant without imposing undue hardship upon the employer (British Columbia Public Service Employee Relations Commission v BCGSEU (1999) 3 S.C.R.3).

Determining the physical and physiological demand of critical tasks

Having completed a task analysis and determined the MOBP and the MPS of the critical tasks, the physical and physiological demands of the task should be quantified (Reilly 2007; Gumieniak et al. 2011; Tipton et al. 2013; Milligan 2013; Taylor et al. 2015*b*). There are a number of factors that can impact the content, criterion, and construct validity and reliability of this stage, which include the following:

1. *Measuring the correct physical and/or physiological attribute of the critical task* — To determine the physical and physiological demand of the critical tasks, one first needs to understand the physiological mechanisms that support the completion of the critical task; e.g., are critical tasks aerobic or anaerobic in nature or a combination of the two.
2. *Choosing the correct sample population* — Data should be collected from a "representative cohort of individuals", which could be a sample from the wider population of those that could apply for a job and those that are currently in the job (Tipton et al. 2013). If a PES comprises a selection test that contains a skilled component that reduces physiological demand, applicants applying for the job should be given sufficient training to obtain this level of skill, as it would be unjustifiable to base selection on an attribute that will be obtained whilst employed (Jackson 1994). If such training is not provided, the "experience" factor should be inherent in the initial assessment of a task but not be used as a rationale for reducing a PES to the level of those most efficient (skilled), as the need for fitness to do a task precedes the opportunity to develop the skill on the task. Therefore, the physiological demand of a task should be mea-

sured on participants with a range of experience levels to determine a minimum acceptable level of performance, and highlight the potential for different PESs for applicants versus incumbents.

3. *Reliability and validity of the equipment used to determine the physical and/or physiological demands* — When measuring the physical and physiological demands of the critical tasks it is important that the measurement tools are valid and reliable. The accuracy of all equipment used in the development of a PES should be reported and calibrated pre- and post-measurements being taken.
4. *Impact of the environment* — The environment in which the measurements are taken needs to be standardized in terms of conditions. For example, dimensions and accessibility of the work space; posture; terrain; protective clothing; urgency of the task; temperature, humidity; location (e.g., indoors/outdoors) (Taylor and Groeller 2003). Failure to do so may invalidate any data collected.

Testing options for PESs

The tests that constitute a PES must either be an accurate direct simulation of the critical task, undertaken using the MOBP, at the MPS (Reilly et al. 2006*b*; Reilly 2007), or, if it is not possible to use simulation, a simple-to-measure test that adequately predicts performance of the critical tasks can be employed (PST). A PST can be developed as long as they have been derived from tests that accurately predict performance on the critical tasks; it is then the criterion predictive validity and reliability of these test scores that are evaluated (Society of Industrial and Organizational Psychology Inc. 2004).

The employee selection criteria used in the past by organizations have been challenged in court and deemed discriminatory because the PST chosen to assess suitability for the job failed to demonstrate how selection related to the job (British Columbia Public Service Employee Relations Commission v BCGSEU (1999) 3 S.C.R.3; Shephard and Bonneau 2002). A PST must be both reliable and valid, as a test cannot be considered valid if it is not reliable; i.e., if successive trials do not yield the same results then the test cannot be trusted. On the other hand, a test can be reliable yet not valid (reproducibly wrong).

Many statistical tests have been proposed for the evaluation of criterion and construct validity and reliability of PSTs. The most common methods involve the use of correlation coefficients and/or the use of difference tests (*t* tests, ANOVAs) (Atkinson and Nevill 1998). However, a strong correlation between test and retest can exist if a systematic improvement (familiarization or training) occurs, which will tell you nothing about the actual reliability of the test (Petersen et al. 2010). Whilst *t* tests and ANOVA can detect a statistical difference between means and therefore detect large systematic bias, no information is provided about random variation (Atkinson and Nevill 1998). Other methods include the use of the coefficient of variation (CV), standard error of the mean (SEM), and limits of agreement (Bland and Altman 1986; Atkinson and Nevill 1998; Petersen et al. 2010). The CV (CV = (SD/mean) × 100) is a dimensionless statistic that permits reliability assessment between different measurement tools. Values are often arbitrarily set at 10% or below, which assumes that 68% of the difference between tests lies within 10% of the mean of the data (Strike 1991); these tests should therefore be used with caution or in combination with other methods. The SEM (SEM = SD/$\sqrt{N}$) is a numeric value that indicates the amount of error that may occur when a random sample mean is used as a predictor of the mean of the population from which it was drawn. The general consensus is that the lower the SEM the more reliable a test (Vincent and Weir 2005). However, how these values should be interpreted in the context of a PES remains unaddressed. Bland and Altman plots in combination with 95% limits of agreement are becoming the preferred method of reliability analysis as they can be used to judge whether changes in performance are real or measurement error, and whether substitute methods are sufficiently reliable or not (Bland and Altman 1986; Atkinson and Nevill 1998).

To establish the implication of measurement error and the criterion and construct validity and reliability of a PST it is recommended that a number of statistical methods are used (Atkinson and Nevill 1998). A reasonable way to deal with the inherent error present with any prediction is to create "pass", "borderline", and "fail" categories obtained from the 75% and 99% prediction intervals, as opposed to confidence intervals (Tipton et al. 2013). The relative likelihood of "false positives" and "negatives" in each of these categories then determines the action taken within each.

In some countries (e.g., USA), this choice does not exist as PSTs for the assessment of occupational performance are prohibited pre-employment. Whilst DTSs are considered by many to be the most valid form of assessment (Williams-Bell et al. 2009; Jamnik et al. 2010) and tend to be favoured by the courts (Berkman v. City of New York 1983; Henderson et al. 2007), this method of assessment cannot identify the percentage of maximal effort that individuals are working at to achieve a pass. They may be working close to their maximum effort to complete an essential task to the MPS. Alternatively, PSTs are generally less time-consuming, can be performed in a controlled environment, and in some cases (e.g., the Tecumseh step test) can be low impact and therefore reduce the risk of individuals suffering injury whilst undertaking a PES. There are many considerations for an employer and the research team before deciding whether to employ a DTS or a PST for the PES.

The arguments for and against the use of DTSs and PSTs are not new. Whilst the courts tend to prefer the use of DTSs (Berkman v. City of New York 1983; Henderson et al. 2007) because of the content and logical validity of these tests, a PST does often allow more individuals to be tested in less time, with a lower risk of injury (Arnold et al. 1982; Thomas and Nelson 2001) and reduced skill component. One option is to use an integrated approach, where PSTs are used alongside DTSs. One organization that has already adopted a similar approach is the United Kingdom Royal National Lifeboat Institution (RNLI); in their standards for RNLI boat crew (Reilly 2007), individuals can only undertake some of the proposed DTSs if they have achieved a pass on a PST that indicates they will not be injured or working maximally on the DTS. In other standards, such as the United Kingdom Maritime and Coastguard Agency's PES for Her Majesty's Coastguards, the use of DTS is either performed after a PST or incrementally, starting with a lower load, to reduce the risk of injury (Milligan 2013). This approach combines consideration of the health and safety of individuals undertaking the fitness standard, with the maintenance of a high level of logical validity.

Setting cut-scores

The "arbitrariness" of setting cut-scores should be considered (Kane 1994; Zumbo 2016). This sense of arbitrariness is reinforced by the oft-heard remark that standard-setting and hence setting a cut-score is, in its essence, a policy decision. Kane (1994, p. 426) states: "there is an element of judgment in all standard setting which is arbitrary in the sense that there is a range of legitimate choices that could be made, but standards vary in their arbitrariness." This has direct links to the setting of the MPS discussed earlier, whereby the weaker the levels of evidence used the more arbitrary the standard-setting will be. Those standards that use what are considered to be high levels of evidence do not seem to be arbitrary; e.g., the minimum time a beach lifeguard should be able to paddle 300 m, or swim 200 m, is based on the time taken for a casualty to drown and the likelihood of a successful resuscitation based on published case studies (Reilly et al. 2006*a*; Table 3). This standard does not seem particularly arbitrary because it is derived from physiological principles. For greater clarity, acknowledging that setting a cut-score has a certain amount of arbitrari-

ness is to acknowledge that when one accepts a cut-score, one is also accepting a certain amount of (minimal) potential classification error. The potential misclassification may come from measurement error on the test score and from what is referred to as construct irrelevant variance or construct underrepresentation in terms of the domain tested. For example, construct irrelevant variance may arise from task easiness or difficulty, which can be traced to equipment design or coaching while measuring the physical abilities construct of focus in the test (Messick 1989; Kane 2006; Zumbo 2007). For a detailed discussion on how to set a valid cut-score, please refer to Zumbo (2016).

Quantifying sources of measurement error (reliability)

To ensure minimal measurement error (reliability), the systematic (e.g., the general learning or fatigue whilst performing a selection test) and the random error (e.g., the biological or mechanical variation of applicants and/or incumbents undertaking the tests or those administering the tests) must be taken into consideration (Atkinson and Nevill 1998; Boyd et al. 2015). It is important in the development of a PES to meaningfully quantify both systematic and random error of the selection tests to ensure that these tests are effective for practical use as the statistical significance (Atkinson and Nevill 1998). For those designing and implementing PESs the consequences of this error should be understood. This form of analyses is often overlooked, but provides practical information for both the researchers and organizations implementing a PES (Spiering et al. 2012).

When conducting a selection test, there is a certain degree of biological variation (e.g., time of day, sleep, fatigue, nutrition, and hydration) (Coulson and Archer 2011; Boyd et al. 2014). It has been suggested that biological variation should form a large proportion of random error with the responsibility falling to the test administrators to minimize technical and environmental variability (e.g., equipment calibration, test circuit set-up) (Boyd et al. 2014). If all else is controlled for (e.g., familiarization) without causing a training effect, then the remaining error can be attributed to biological variation and could result in the implementation of a zone around the cut-score where test scores are considered inclusive (Boyd et al. 2014), further supporting the inclusion of a borderline category to take into account the inherent error within any test (Tipton et al. 2013).

A problem for researchers designing PES assessments is assessing the reliability (amount of measurement error) in the resultant test score. For example, a test–retest study can be conducted; however, that study, on its own, ignores variation in test scores because of the mediators and moderators of task performance, such as task difficulty, equipment, number of tasks, load, and other sources of unreliability or inconsistency within a complex assessment. Generalizability theory allows the use of ANOVA-type procedures in which the variability in numerous factors can be estimated simultaneously. The relative proportion of variation attributed to these various factors in form of reliability coefficients (sometimes called generalizability coefficients) can then be calculated. Generalizability theory is intended to pinpoint the sources of measurement error, disentangle them, and estimate each one (Cronbach et al. 1972; Shavelson and Webb 1981, 1991; Brennan 2001).

Several recent studies have used an ANOVA framework and the intra-class correlation statistic so there is familiarity in the PES field with these statistics (Payne and Harvey 2010; Burnstein Steele and Shrier 2011; Spiering et al. 2012; Boyd et al. 2015). For example, in a recent study the variability in performance on a DTS test of physical fitness for firefighters was investigated (Boyd et al. 2015). These researchers focused on practice, pacing strategy, and day-to-day fluctuations in biological function. Their goal was to investigate what proportion of test score variation can be attributable to some of these sources independently. Payne and Harvey (2010) also draw attention to sources of test-score variation by investigating test–retest and sources of measurement error, such as task demands, separately. Spiering et al. (2012) assessed the reliability of 7 military relevant occupational tests; the results showed that 3 of the tests required familiarization before a stable value could be obtained. Boyd et al. (2015) conducted 6 trials of the firefighters circuit and found that although only small improvements were seen at the sixth trial, they were still statistically significant. In fact of 51 subjects, 15 achieved their best score on the fifth test and 32 subjects achieved their best score on the sixth test. However, none of these studies have taken it a step further to conduct generalizability theory analysis. That is once the various sources of measurement error have been quantified, "what-if" calculations, called D-studies, can be implemented to investigate the gain in reliability; e.g., the need for familiarization. In short, generalizability theory considers assessment akin to an experiment wherein the proportion of variation can be calculated by using a statistical test such as intra-class correlation coefficients (ICCs) for various factors in the assessment design. For detailed step by step guidelines, readers are guided to Bloch and Norman (2012) and Briesch et al. (2014).

To put generalizability theory into the context of a PES, the reader is asked to consider assessing the validity and the reliability of a test (e.g., a timed circuit) to be used in a PES. By using a repeated-measures ANOVA design, 3 hypotheses can be tested and the concept of generalizability theory can be applied. First, the validity of the selection test can be tested against the "gold standard". Second, the reliability of the selection test can be measured across repeated trials, and finally the interaction effect between selection test and trials can be determined, whereby the researcher does not only consider the test–retest or task variation in isolation, but can study these factors together, providing a sense of relative magnitude as well as gathering information on the interaction. This assumes that each factor in the test has a statistical distribution and hence a variance. In generalizability theory the variance component for the testee (e.g., the incumbent) is called the universe-score variance. The variance components for the other factors are considered error variation, and each variance component can be estimated using traditional ANOVAs or methods such as Maximum Likelihood.

The relative magnitudes of the estimated variance components are calculated in the form of ICCs and provide information about sources of error influencing the PES test performance. In addition, standard errors for variance components provide information about sampling variability. Finally, generalizability theory allows tests to have decisions assisted by norm-referencing or criterion-referencing (Zumbo 2016). These 2 options reflect 2 types of definitions of error variance and require different types of ICCs and reliability coefficients (Bloch and Norman 2012; Briesch et al. 2014). Whilst the use of ICCs has been advocated, Atkinson and Nevill (1998) have recommended that they should not be the sole statistic used, as more work is needed to define acceptable ICCs based on the realisation of definite analytical goals.

Conclusion

Validation is about presenting empirical evidence and a compelling argument to support the intended inference and to show that alternative or competing inferences are not more viable. In particular, the aim is to identify the degree to which construct under-representation and construct-irrelevant variance are problems. Because of changes occurring in time (e.g. in empirical knowledge, theoretical understandings, or values, society), the process of validation is an on-going one. The on-going nature is particularly relevant to MPSs and deriving a cut-score, which clearly are not set once and for always, but need to be revisited as the key elements of the validity argument change, including societal expectations for job performance.

A defensible PES should be valid and reliable, using an evidence-based approach including a standardized and logical method to progress from task analysis to PES. A central part of test validation then consists of a demonstration that the proposed pass score can be interpreted as representing an appropriate MPS. An important component in this process is standard-setting through the establishment of the minimum acceptable pace/intensity (MPS) for the completion of critical tasks; this determines the final pass/fail criteria for a PES. The establishment of an MPS often involves input from SMEs, chosen in an unsystematic way, and whose decisions are based on what would be regarded as weak evidence in other disciplines. Setting a cut-score without a description of the conceptual version of the desired levels of competence (the performance standard) or of any evidence external or internal to the test data results in a cut-score that is capricious and indefensible. Given the ultimate potentially significant consequences of PES, the level of evidence of the methods used to develop them should be given a great deal of consideration.

References

AERA. 1999. Standards for educational and psychological testing. American Psychological Association, Washington, DC, USA.

AERA. 2014. Standards for Educational and Psychological Testing. American Psychological Association, Washington, DC, USA.

Arnold, J., Rauschenberger, J., Soubel, W., and Guion, R. 1982. Validation and utility of a strength test for selecting steelworkers. J. Appl. Psychol. **67**(5): 588–604. doi:10.1037/0021-9010.67.5.588.

Atkinson, G., and Nevill, A.M. 1998. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. **26**(4): 217–238. doi:10.2165/00007256-199826040-00002. PMID:9820922.

Berkman v. City of New York, 705 F.2d 584 (2d Cir. 1983).

Bilzon, J.L.J., Scarpello, E.G., Bilzon, J.L.J., and Allsopp, A.J. 2002. Generic task-related occupational requirements for Royal Naval personnel. Occup. Med. **52**(8): 503–510. doi:10.1093/occmed/52.8.503.

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.

Bland, J.B., and Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, **1**: 307–310. PMID: 2868172.

Bloch, R., and Norman, G. 2012. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Med. Teach. **34**(11): 960–992. doi:10.3109/0142159X.2012.703791. PMID:23140303.

Bonneau, J. 2001. Evaluating physical competencies fitness related tests task simulation or hybrid. Objectives, process and consensus summary of the national forum on bona fide occupational requirements. *In* Proceedings of the Consensus Forum on Establishing Bona Fide Requirements for Physically Demanding Occupations. *Edited by* N. Gledhill, J. Bonneau, and A. Salmon. York University, Toronto, Ont., Canada. pp. 23–33.

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

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**: 364–370. doi:10.1139/apnm-2014-0281. PMID:25781347.

Brennan, R.L. 2001. Generalizability Theory. Springer–Verlag, New York, N.Y., USA.

Briesch, A.M., Swaminathan, H., Welsh, M., and Chafouleas, S.M. 2014. Generalizability theory: a practical guide to study design, implementation, and interpretation. J. School Psychol. **52**: 13–35. doi:10.1016/j.jsp.2013.11.008.

British Columbia Public Service Employee Relations Commission v. BCGSEU, 3 S.C.R. 3 Meirion Decision (Internet). [Supreme Court of Canada.] Ottawa, Ontario; Government of Canada; 1999. Available from scc-csc.lexum.com/scc-csc/scc-csc/en/item/1724/index.do. [Accessed January 2013.]

Burnstein, B.D., Steele, R.J., and Shrier, I. 2011. Reliability of fitness tests using methods and time periods common in sport and occupational management. J. Athl. Training, **46**: 505–513. PMID:22488138.

Chatterjee, S., Chatterjee, P., Mukherjee, P.S., and Bandyopadhyay, A. 2004. Validity of Queen's College step test for use with young Indian men. Br. J. Sports Med. **38**: 289–291. doi:10.1136/bjsm.2002.002212. PMID:15155428.

Conn, A.W., Miyasaka, K., Katayama, M., Fujita, M., Orima, H., Barker, G., and Bohn, D. 1995. A canine study of cold water drowning in fresh versus salt water. Crit. Care Med. **23**(12): 2029–2037. doi:10.1097/00003246-199512000-00012. PMID:7497726.

Coulson, M., and Archer, D. 2011. Practical Fitness Testing: Analysis in Exercise and Sport. Bloomsbury Publishing, London, UK.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. 1972. The Dependability of Behavioral Measurements. Wiley, New York, N.Y., USA.

Eccles, M., and Mason, J. 2001. How to develop cost-conscious guidelines. Health Technological Assess. **5**(16): 1–69. PMID:11427188.

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.

Fainer, D.C., Martin, C.G., and Ivy, A.C. 1951. Resuscitation of dogs from fresh water drowning. J. Appl. Physiol. **3**: 417–426. PMID:14803394.

Gledhill, N., Jamnik, V., and Shaw, J. 2001. Establishing a bona fide occupational requirement for physically demanding occupations. *In* Proceedings of the Consensus Forum on Establishing Bona Fide Requirements for Physically Demanding Occupations. *Edited by* N. Gledhill, J. Bonneau, and A. Salmon. York University, Toronto, Ont., Canada. pp. 9–13.

Golden, F., and Tipton, M. 2002. Essentials of Sea Survival. Human Kinetics, Leeds, UK.

Gumieniak, R., Jamnik, V.K., and Gledhill, N. 2011. Physical fitness bona fide occupational requirement for safety-related physically demanding occupations; test development considerations. Health Fitness J. Can. **4**(2): 47–52.

Henderson, N.D., Berry, M.W., and Matic, T. 2007. Field measures of strength and fitness predict firefighter performance on physically demanding tasks. Personnel Psychol. **60**(2): 431–473. doi:10.1111/j.1744-6570.2007.00079.x.

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

Jamnik, V.K., Thomas, S.G., Shaw, J.A., and Gledhill, N. 2010. 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.

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

Kane, M.T. 2006. Validation. *In* Educational Measurement. 4th ed. *Edited by* R.L. Brennan. American Council on Education/Praeger, Westport, Conn., USA. pp. 17–64.

Lavin, R.P., Dreyfus, M., Slepski, L., and Kasper, C.E. 2007. Subject matter experts: facts or fiction. Nursing Forum, **42**(4): 189–195. doi:10.1111/j.1744-6198.2007.00087.x. PMID:17944700.

McArdle, W.A., Katch, F.I., and Katch, V.L. 2007. Exercise physiology – energy, nutrition and human performance. 6th ed. Lippincott Williams and Wilkins, USA.

Messick, S. 1989. Validity. *In* Educational Measurement. 3rd ed. *Edited by* R.L. Linn. American Council on Education and Macmillan, New York, N.Y., USA. pp. 13–103.

Milligan, G.S. 2013. Fitness standards for the Maritime and Coastguard Agency and the oil and gas industry. PhD. thesis, University of Portsmouth, Portsmouth, UK.

Morley, P.T., Atkins, D.L., Billi, J.E., Bossaert, L., Callaway, C.W., de Caen, A.R., et al. 2010. Part 3: Evidence evaluation process: 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation, **122**(16 Suppl. 2): S283–S290. doi:10.1161/CIRCULATIONAHA.110.970947. PMID:20956251.

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

Petersen, A., Payne, W., Phillips, M., Netto, K., Nichols, D., and Aisbett, B. 2010. Validity and relevance of the pack hike wildland firefighter work capacity test: a review. Ergonomics, **53**(10): 1276–1285. doi:10.1080/00140139.2010.513451. PMID:20865610.

Pheasant, S., and Haslegrave, C.M. 2005. Bodyspace: Anthropometry, ergonomics and the design of work. CRC Press.

Phillips, M., Payne, W., Lord, C., Netto, K., Nichols, D., and Aisbett, B. 2012. Identification of physically demanding tasks performed during bushfire suppression by Australian rural firefighters. Appl. Ergon. **43**: 435–441. doi:10.1016/j.apergo.2011.06.018. PMID:21802652.

Quan, L., Mack, C.D., and Schiff, M.A. 2014. Association of water temperature and submersion duration and drowning outcome. Resuscitation, **85**(6): 790–794. doi:10.1016/j.resuscitation.2014.02.024. PMID:24607870.

Rayson, M.P. 1998. The development of physical selection procedures. Phase 1: Job analysis. Contemporary Ergonomics, 1998: 393–397.

Rayson, M.P. 2000. Job analysis. *In* The Process of Physical Fitness Standards Development. State of the Art Report. *Edited by* S. Constable and B. Palmer. Human Systems Information Analysis Center, US Department of Defense, Washington, DC, USA. pp. 67–100.

Reilly, R., Zedeck, S., and Tenopyr, M. 1979. Validity and fairness of physical ability tests for predicting performance in craft jobs. Appl. Psychol. **64**(3): 262–274. doi:10.1037/0021-9010.64.3.262.

Reilly, T. 2007. Fitness Standards for the Royal National Life Boat Institution (RNLI) Lifeboat Crew. PhD. dissertation, University of Portsmouth, Portsmouth, UK.

Reilly, T., and Tipton, M. 2005. Task-based standards for lifeboat crew: avoiding ageism. International Congress Series, **1280**: 219–223. doi:10.1016/j.ics.2005.02.078.

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**: 6-11. doi: 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**: 12–17. doi:10.1093/occmed/kqi168.

Rogers, T.W., Docherty, D., and Petersen, S. 2014. Establishment of performance standards and a 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.

Shavelson, R.J., and Webb, N.M. 1981. Generalizability theory: 1973–1980. Br. J. Math. Stat. Psychol. **34**: 133–166. doi:10.1111/j.2044-8317.1981.tb00625.x.

Shavelson, R.J., and Webb, N.M. 1991. Generalizability Theory: A Primer. Sage Publications, Newbury Park, Calif., USA.

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. PMID:12180318.

Siconolfi, S.F., Garber, C.E., Lasater, T.M., and Carleton, R.A. 1985. A simple, valid step test for estimating maximal oxygen uptake in epidemiologic studies. Am. J. Epidemiol. **121**(3): 382–390. PMID:4014128.

Siddall, A.G., Standage, M., Stokes, K.A., and Bilzon, J.L.J. 2014. Development of Occupational Fitness Standards for the UK Fire and Rescue Services (FRS). Available from firefitsteeringgroup.co.uk/UKFRS_Fitness_Standards_Final_Report.pdf. [Accessed 2015.]

Sireci, S.G. 1998. The construct of content validity. Soc. Indic. Res. **45**(1): 83–117. doi:10.1023/A:1006985528729.

Society for Industrial and Organizational Psychology Inc. 2004. Principles for the Validation and use of Personnel Selection Procedures. 4th ed. pp. 1–76. siop.org/_Principles/principles.pdf. [Accessed 2010.]

Spiering, B.A., Walker, L.A., Hendrickson, N.R., Simpson, K., Harman, E.A., Allison, S.C., and Sharp, M.A. 2012. Reliability of military-relevant tests designed to assess soldier readiness for occupational and combat-related duties. Mil. Med. **177**: 663–668. doi:10.7205/MILMED-D-12-00039. PMID:22730841.

Strike, P.W. 1991. Statistical Methods in Laboratory Medicine. Butterworth–Heinemann, Oxford, UK.

Sykes, K., and Roberts, A. 2004. The Chester step test – a simple yet effective tool for the prediction of aerobic capacity. Physiotherapy, **90**: 183–188. doi:10.1016/j.physio.2004.03.008.

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.

Taylor, N.A., Fullagar, H.H., Mott, B.J., Sampson, J.A., and Groeller, H. 2015*a*. 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., Fullagar, H.H., Sampson, J.A., Notley, S.R., Durley, S.D., Lee, D.S., and Groeller, H. 2015*b*. 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, R. 1990. Interpretation of the correlation coefficient: a basic review. J. Diagn. Med. Sonog. **1**: 35–39.

Thomas, J.R., and Nelson, J.K. 2001. Research Methods in Physical Activity. 4th ed. Human Kinetics.

Thomas, J.R., Nelson, J.K., Silverman, S., and Silverman, S.J. 2005. Research Methods in Physical Activity. 5th ed. Human Kinetics.

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.

Vincent, W.J., and Weir, J.P. 2005. Statistics in Kinesiology. 3rd ed. Human Kinetics.

Williams-Bell, M.F., Villar, R., Sharratt, M.T., and Hughson, R.L. 2009. Physiological demands of the firefighter candidate physical ability test. Med. Sci. Sports Exerc. **41**(3): 653–662. doi:10.1249/MSS.0b013e31818ad117. PMID:19204584.

Zumbo, B.D. 2007. Validity: Foundational issues and statistical methodology. *In* Handbook of Statistics: Psychometrics. Vol. 26. *Edited by* C.R. Rao and S. Sinharay. Elsevier, Amsterdam, Netherlands. pp. 45–79.

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

Zumbo, B.D., and Rupp, A.A. 2004. Responsible modelling of measurement data for appropriate inferences: important advances in reliability and validity theory. *In* The SAGE Handbook of Quantitative Methodology for the Social Sciences. *Edited by* D. Kaplan. Sage Press, Thousand Oaks, Calif., USA. pp. 73–92.