Applied Physiology, Nutrition, and Metabolism
Physiologie appliquée, nutrition et métabolisme

Standard-setting methodology: Establishing performance standards and setting cut scores to assist score interpretation

Journal:	<i>Applied Physiology, Nutrition, and Metabolism</i>
Manuscript ID	apnm-2015-0522.R2
Manuscript Type:	Review
Date Submitted by the Author:	15-Feb-2016
Complete List of Authors:	Zumbo, Bruno; University of British Columbia, ECPS
Keyword:	standard setting, physical fitness test, cut score, pass score, test validity

Standard-setting methodology: Establishing performance standards and setting cut scores to assist score interpretation

Bruno D. Zumbo
University of British Columbia

Citation:

Zumbo, B.D. (in press). Standard-setting methodology: Establishing performance standards and setting cut scores to assist score interpretation. *Applied Physiology, Nutrition, and Metabolism*.

Author Note:

Prepared by invitation for the: ***Second International Conference on Physical Employment Standards***, Canmore, Alberta, August 2015. Bruno D. Zumbo is Paragon UBC Professor of Psychometrics and Measurement. He would like to thank Dr. Amery Wu, Dr. Anne Gadermann, and Michelle (Yue) Chen for feedback on an earlier version of this paper.

Address correspondence to: Bruno Zumbo, Paragon UBC Professor of Psychometrics & Measurement, University of British Columbia, Dept. of ECPS, Scarfe Building, 2125 Main Mall, Vancouver, B.C. CANADA V6T 1Z4 e-mail: bruno.zumbo@ubc.ca

Abstract

A critical step in the development and use of tests of physical fitness for employment purposes (e.g., fitness for duty) is to establish one or more cut points dividing the test score range into two or more ordered categories reflecting, for example, fail/pass decisions. Over the last three decades elaborated theories and methods have evolved focusing on the process of establishing one or more cut scores on a test. This elaborated process is widely referred to as ‘standard setting’. As such, the validity of the test score interpretation hinges on the standard setting, which embodies the purpose and rules according to which the test results are interpreted. The purpose of this paper is to provide an overview of standard setting methodology. The essential features, key definitions and concepts, and various novel methods of informing standard-setting will be described. The focus is on foundational issues with an eye toward informing best practices with new methodology. Throughout, a case is made that in terms of best practices, establishing a test standard involves, in good part, setting a cut score and can be conceptualized as evidence/data-based policy making that is essentially tied to test validity and an evidential trail.

Keywords: Standard setting, physical fitness testing, cut score, pass score, test validity, performance standards, psychometrics

Tests of physical fitness for employment purposes (e.g., fitness for duty) are widely used in a variety of settings, including emergency response occupations such as law enforcement or firefighting, and a variety of roles in military service. These test results are nearly always used to make pass/fail decisions. Over the last three decades elaborated theories and methods have evolved in the discipline of testing and measurement focusing on the process of establishing one or more cut scores on a test. This elaborated process is widely referred to as ‘standard setting’, which embodies the purpose and rules according to which the test results are interpreted (see, for example, Cizek, 2012; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006).

The purpose of this paper is to provide a type of knowledge translation and reflection on the state of the art of standard setting methodology for testing of physical fitness for employment purposes from the perspective of developments in the discipline of testing and assessment, broadly defined. Developments in standard setting methodology have largely arisen in educational, language, psychological, and health assessment, licensure and certification testing, and medical education. It should be noted that because of the voluminous measurement and assessment literature that has accumulated over the last 30 years, throughout this paper the focus is on foundational issues and broad overview with an eye toward informing best practices and describing new developments. The paper is divided into five sections: (1) definitions and concepts, (2) a brief overview of methods for setting cut-scores with an eye towards new methodological developments, (3) the importance of conditional estimates of precision in test scores, (4) how standards are tied to test validation and a focus on test consequences, and (5) closing remarks.

Definitions and Concepts

In a seminal paper, Kane (1994) cleared a great deal of confusion in the standard setting literature by articulating a distinction between performance standards and setting cut-scores. By drawing a connection to test validation, Kane showed how it is useful to draw a distinction between (i) the cut-score (which he refers to as a ‘passing score’), defined as a point on the test score scale, and (ii) the performance standard, defined as the minimally adequate level of performance for some purpose. In describing this distinction, as motivated by test validation, he writes:

“Validation then consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard. The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version of the desired level of competence. In much of the literature on standard setting, the distinction between the passing score and the performance standard is not explicitly drawn, making it difficult to evaluate the validity of the interpretation assigned to the passing score. Maintaining a clear distinction between the passing score and the corresponding performance standard helps to avoid much confusion.” (Kane, 1994, pp. 425-426).

Although this distinction is not yet universally made in the standard setting literature (Hambleton, 2001; Kane, 2001; Rogers and Ricker, 2006), Rogers et al. (2014) demonstrated that this distinction between the performance standard and the cut-score is particularly useful in the domain of tests of physical fitness for employment purposes. Rogers et al. reported on the standard setting for the Canadian Forces Firefighter Physical Fitness Maintenance Evaluation (FF PFME), a test that is comprised of 10 physical tasks where the test score is the overall completion time – please see Petersen et al. (2013) for details. Based on the description and definition of the variable they are measuring, and the context and purpose of the test, these authors made the case for three levels of performance: acceptable work capacity, minimally acceptable work capacity, and unacceptable work capacity. Table 2 of Rogers et al. provides a

description of these three levels of performance. For example, acceptable work capacity is characterized as:

“Safe – working in an efficient manner – not too fast or too slow – and in a manner that does not endanger the firefighter, co-workers or the public (i.e. reflecting a duty of care for co-workers, public and property). Effective – performed with a sense of purpose consistent with accomplishing first-response firefighting activities in an efficient way. Does not require intervention by a supervisor.” (p. 1753).

In short, ‘standards’ should help understand proficiency and performance at a particular level and how those levels concatenate so as to inform setting cut scores and report back to the test-taker and stake-holders. The standards should inform the test taker about what they can do and what they can improve upon.

Performance standards usually contain performance level labels such as ‘pass’ or ‘fail’ for two categories, or ‘beginning’, ‘intermediate’ and ‘advanced’ if there more than two categories. To be useful and effective in the standard setting process, performance standards must include more than just performance labels; instead standards need to include longer elaborations that are intended to more fully describe what the label (e.g., ‘pass’) indicates about what a test taker can do. A clear description of what a test taker can do goes a long way to helping establish defensible cut-scores. The standard setting study reported in Rogers et al. (2014) is a good example of how clearly stated standards were then used to inform the process of setting cut scores and supporting the defensibility of the cut scores.

With the description of the conceptual version of the desired levels of competence, i.e., the performance standard, at hand the next step is to translate these descriptions to the operational version of the desired level of competence by setting the cut-score on the test. In so

doing, the act of setting the cut-score is less likely to be confounded by the yet undeclared conceptual version of the desired competence. That is, there is clarity of purpose that is provided by distinguishing between the performance standard from the cut-score; making the evidential trail of how one sets the cut-score more defensible. This is not unlike the direction and clarity of purpose provided by having first fully articulated the domain of interest (or construct) as well as the purpose of the test before test design and test construction.

At this point, it is important to remind ourselves that an examiner's objective is to form an accurate decision about the true underlying state of the test-taker, based on the observed score. Standard setting involves finding the point (cut-score) on the distribution of total test scores that meaningfully categorizes examinees into two or more groups reflecting increasing level of performance. Standard-setting procedures used in testing for physical ability reflect maximal performance. As we describe above, in such tests, decisions about examinees are based on specific cut-scores that represent a standard of performance. The cut-score marks the minimum level of performance required for a passing score. Standards can be set using either a norm-referenced or a criterion-referenced framework; this is described in more detail in the next section of the paper, or some hybrid of the two.

In addition to clarifying the difference between performance standards and setting cut-scores, Kane (1994) went on to address what Glass (1978) and others have described as the 'arbitrariness' of setting cut-scores. 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. As Kane notes, "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" (p. 426). He goes on to say that indeed some standards seem quite arbitrary. Kane

describes a tradition that many of us have experienced in our educational backgrounds; a teacher or school administrator declares that a pass on a test requires a minimum test score of 70% correct, with no other justification for the 70% cut-score except, for example, tradition or school history. In most applications like this there is no justification for the pass-score of 70; why not 68 or 73? In addition, making such a pass-score depends on the items (easy tests could result in all test takers passing the 70% score on the test; whereas difficult tests could result in no one passing the 70% mark). The inherent problem with having a cut-score like the one described is that it is driven by neither the test performance of a normative test-taker sample, nor the content or tasks in the test.

Some standards, on the other hand, do not seem at all arbitrary; Kane uses the example of a test requirement that a lifeguard be able to swim a certain distance in a certain time and then swim back pulling a struggling victim. This standard does not seem particularly arbitrary because it is likely derived from information given in a job or task analysis and an understanding of what it means to be a lifeguard.

For greater clarity, acknowledging that setting a cut-score has a certain amount of arbitrariness is to acknowledge that when one accepts a cut-score, one is also accepting a certain amount of (minimal) potential classification error rate. The potential misclassification may come from measurement error on the test score, and what is referred to as construct irrelevant variance or construct underrepresentation in terms of the domain one is testing. For example, construct irrelevant variance may arise from irrelevant task easiness or difficulty that can be traced to equipment design or coaching while measuring the construct of focus of the test, in this case the physical abilities (Kane, 2006; Messick, 1989; Zumbo, 2007).

One also needs to acknowledge that accepting a cut-score involves a balance of costs (financial or human resource), as well as a desire to be more or less inclusive. This is best captured by the methods of decision theory (sensitivity, specificity, positive predictive value, and negative predictive value) and signal detection theory. Zumbo et al. (2002) provide an overview of these concepts in the context of test validation and the use of tests for screening, selection, and assessment. Sensitivity, specificity, and predictive values are all used as evidence of the accuracy or correctness of a decision (i.e., validity) to categorize individuals – in this case, as pass or fail. Before one can calculate these values, one must determine by some means other than the test, for example a gold standard, who in the sample passed and who failed. Therefore, if one knows who should pass and who should fail, based on their performance on the gold standard, one can identify false-positive errors (test takers who pass the test but have a failing score on the gold standard) and false-negative errors (test takers who fail the test, but have a passing score on the gold standard). Following this same reasoning, true negative and true positives are test takers who match in terms of their classification on the test and the gold standard. These classifications can be computed for each possible passing score. Figure 1 is a depiction of a two by two, row by column, table where a positive event is a pass, and a negative event a fail and the various decision theoretic outcomes are defined.

Insert Figure 1 about here

In terms of testing for physical fitness, like all testing, the balance of errors depends greatly on the testing purposes: selection, screening, or diagnosis; as well as, the test taker

population such as new recruit or incumbent. For example, sensitivity and specificity tend to be of greater interest to the researcher who is interested in the accuracy of a test (and its corresponding performance standard and cut-score) in identifying both those individuals who are physically fit for duty and those individuals not physically fit for duty. On the other hand, predictive values may be of greater interest to administrators (e.g., fire chiefs), test takers, and individuals concerned with litigation because they indicate how accurately the test (and its corresponding performance standard and cut-score) can predict the presence or absence of physical fitness for duty when it is not known whether the person is physically fit for duty. As Zumbo et al. (2002) note, the test purpose and test population help determine which decision-theoretic statistics should be of focus.

To this point we have highlighted: (a) the distinction between performance standards and setting cut-scores, and (b) that there is an element of judgment in all standard setting which, as Kane states, is arbitrary in the sense that there is a range of legitimate choices that could be made, but standards vary in their arbitrariness. It is important to acknowledge that establishing a test standard is best conceptualized as evidence- and data-based policy making that is essentially tied to test validity. The central concern is having a defensible rationale for describing and adopting a performance standard and then setting the cut-score. A central part of test validation then consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard. 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. The next section of this paper provides a brief overview of methods for gathering data to set a cut-score.

Brief Overview of Methods for Setting Cut-Scores With an Eye Towards New Methodological Developments

The commonly used methods for setting cut-scores have been clearly described in a number of reviews and entire textbooks are devoted to the topic (e.g., Cizek, 2012; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006; Kane, 1998, 2001; McKinley, et al. 2005; Zieky, 2001) and therefore do not need to be described in detail here. Please see Tipton et al. (2013) as a starting point for more specific understanding of the processes involved in standard setting in physiological employment standards. Later in this section, brief descriptions of the most commonly used methods in testing for physical fitness for duty are provided and their strengths, limitations, and methods for improving their defensibility are discussed. Furthermore, some methodological developments will be highlighted.

One can categorize the methods for setting cut-scores a number of different ways (e.g., Cizek, 2012; Hambleton and Pitoniak, 2006; Rogers and Ricker, 2006), but with a broad scope of practice of physical fitness testing in mind, it is useful to classify the various methods as: (a) methods based mainly on the statistical distribution of test scores, or (b) methods based mainly on the judgements of experiential or subject matter experts as raters. Although we focus on the two more widely used classes of methods (statistical distribution or judgement based) separately, one can imagine (and one will see in the descriptions below) that most implementations of these methods attempt to borrow strengths from both classes of approaches in a hybrid strategy.

Methods that are mainly based on the statistical distribution of test scores

Methods based on the statistical distribution of test scores are commonly tied to norm-referenced test interpretation. A norm-referenced interpretation is made when individual test performance is described with respect to some normative sample. For the most part, norm-

referenced approaches are best suited for tests whose purpose is to rank test takers based on highest to lowest test performance. Norm-reference based standard setting methods are traditionally and widely used in setting cut-scores for physical fitness (Jamnik, et al. 2013). As a hypothetical example, one could set the cut-score (and, implicitly define the standard of performance) by simply failing the bottom 10% of a normative sample of test takers. This is provided as an example and not as a recommended approach because, on its own, it lacks defensibility (why 10% and not some other percentage) and is often perceived as problematic because irrespective of the strength of a test-taker sample 10%, in this example, will always not pass.

A method that is used in the physical fitness testing literature is to collect data from a sample of test takers who are currently in an occupational group (e.g., forest firefighter) and calculate the mean or the quantiles (e.g., 25th percentile, the median, 75th, or 95th percentile) of the distribution of test scores (time to complete the circuit). From this information and some measure of variability or spread (such as the standard deviation or standard error) a cut-score is set. These statistics, on their own as the sole evidence for the cut-score, are sometimes described as arbitrary in the standard setting literature because of the fact that the test score, in this case time to complete a circuit, is be considered a function of the level of physical fitness. If it is a continuous function then, on its own without any other external data or evidence, there is no simple and obvious way to choose a particular time as the cut-score. So, the choice of the cut-score is arbitrary in the sense that there is no compelling reason why it could not be set a little higher or a little lower – e.g., why not the 73rd percentile, or any other point on the score distribution. If the function is discontinuous with either a gap or jump in the curve relating time to complete a circuit to the level of physical fitness, then that information can be very helpful in

determining a cut-score – so long as the discontinuity is not an artifact of testing due to equipment or some other variable not of interest in the testing.

In addition, a key strong assumption is made of the normative sample: that just because someone is a member of an occupational group, and hence a member of the normative sample, that they are capable of performing the job adequately. Therefore, evidence needs to be provided that supports the (often implicit) claim that the mean (or any other statistic such as the 90th percentile) of the scores of a normative sample comprised of an occupational group are in fact capable of performing the job adequately – and it is not sufficient to assume so just because they are already in the job.

In addition to supporting the assumption that members of a normative sample can perform a job adequately, there are two sets of methods for increasing the defensibility of this norm-referenced approach. The first and most obvious way to increase the defensibility is to consider the representativeness of the sample of test takers in terms of a target population of test takers which reflect the test purpose. One then imagines sampled and unsampled test takers and one asks whether test takers in these two groups are exchangeable. One, in essence, asks the question of whether the statistical cut score would be the same if one had the unsampled individuals rather than those who are in the normative sample (Zumbo, 2007). The more exchangeable those two sets of individuals, the stronger the inference and conclusions about the statistical decisions from the cut score computed from the sample of test takers – this is the conceptual foundation of Zumbo's Draper-Lindley-de Finetti, DLD, framework (Zumbo, 2007, pp. 56-63) described in more detail below.

The second way one can increase the defensibility of the norm-reference approach is to tie the setting of the cut-score to either an external criterion such as performance on the job, the

results of a job analysis, or a reference group who take the test but are already performing the job. Doing so reduces the fluctuation due to examinee ability or test difficulty by supplementing the test score data with other information. A hybrid of the norm-referenced and criterion-referenced methods can be created by adapting a version of the so-called borderline-group method (Livingston and Zieky, 1982) to help inform setting the cut-score. For example, based on experiential or subject-matter experts' nominations one may be able to identify individuals in your normative sample as borderline, in the sense that their level of physical fitness is right around the performance standard. These nominated individuals would then take the test and their test times could be used, along with the statistics from the normative sample, to set a cut-score. Livingston and Zieky provide several validity checks and recommendations for best practice.

By asking questions of the exchangeability of the normative sample and about the supplemental information one can then begin to consider the inference limits (bounds) from measures. Imagine that the normative sample is comprised of 40 female firefighters in the derivation of the cut-score (Gledhill and Jamnik, 2011, p. 45) then the inferential strength of a cut-score may be moderate-to-strong for one population (e.g., female firefighters) but not another population (e.g., male firefighters). It should be noted that by 'inferential strength' we mean the amount of support that the evidence or reasons provide the conclusion; and is therefore considered a matter of degree such that the more support (the more evidence or reasons) there is for a conclusion, the stronger the argument for the conclusion. At the very least, this framework suggests that one needs to investigate this question of exchangeability and perhaps add further supplemental information to boost the inferential strength. In this case, the framework also raises the provocative question of the inferential strength of a cut-score based on a small sub-sample that is then meant to be applied to all test-takers. Doing so is a policy decision but the framework

asks whether this policy may, in fact, be limiting the inferential strength of the cut-score and hence eventually the test decisions themselves. Again, the framework would suggest that further supplemental information needs to be added such as more information from the job analysis, or stronger explanatory understanding for task performance by male and female firefighters. Zumbo (2007, 2009) presents a theory of validation that focuses on explanatory modeling.

Zumbo (2007) introduces a framework for reasoning about test data (the Draper-Lindley-de Finetti, DLD, framework) that helps clarify some of the issues in test development. For example, it is sometimes stated that job simulation tasks are used so that one does not have to make any inferences from the test scores – performance on the test that simulates a job is purely descriptive and the validity of test inferences are by fiat. Clearly, from the above framework one is still making inferences in “simulation” tests because you still do not have the individual doing the job, *per se*. The simulation may be more realistic hence shortening the inferential distance from test to real job performance, but an inference is still being made from those job simulation tasks to other tasks like them.

Zumbo's (2001, 2007) DLD framework does not only deal with sampled and unsampled test tasks but also the question of the match and justification between the test tasks and the performance in the job. This provides one way of thinking about the tasks in the test (the different elements of the circuit in the test) because they are a sub-set of a (potentially well-defined) finite population of tasks or tasks from which one or more versions of the test may be constructed by selection of a sample of tasks from this larger set of possible tasks. Obviously, creating all possible tasks would be both economically and practically unfeasible. In some situations one has a clearly articulated domain of tasks (perhaps even articulated in the test specifications or based on a job analysis) so that one would have a very good sense of the

exchangeability of the tasks one has with those that are not in the assembled test. In other cases one has very little information about the domain being measured and hence has little knowledge of the exchangeability of the tasks one has with those in the hypothetical task population (or domain). In this case, the data you have is the information on the sampled tasks in your assessment and the data you wish you had is the corresponding information on the unsampled tasks. In essence this gets at how well one knows the construct or domain being sampled and how well one plans (e.g., task specifications) in the process of constructing the assessment.

In summary, setting the cut-score based only on the statistical distribution of test scores is less defensible than doing so with some additional information or evidence. Recall, that setting a cut-score based on the statistics from the normative test sample, on its own, is not a highly defensible practice. Instead, one wants to supplement the statistics from the normative sample with other (external) information such as test performance of a contrast for borderline group, a strong criterion predictor of job performance, information from the job analysis, information from the description of the performance standard, and/or strong explanatory theory of the mediators and moderators of task and test performance. In short, one is asked to assess the strength of the supplemental information to the statistical standard setting. One is asked to imagine a minimally qualified individual and invoke a mental picture of what levels of performance (in our case, completing a circuit of physical fitness tasks) of a just-barely-passing candidate. Examples of prototypical performance will sometimes help in conjuring an image of a passing test-taker who makes a tolerable number of errors and in distinguishing this individual from one whose performance makes not passing more plausible.

Methods based mainly on the judgements of experiential or subject matter experts as raters

Methods based mainly on the judgements of experiential or subject matter experts as raters are commonly tied to criterion-referenced tests that are designed and interpreted with respect to a defined level of ability, or a performance standard (Hambleton, 1984). Performance testing using criterion-referenced tests involves the selection of a cut-score that accurately categorizes examinees into meaningful groups based on specific, defined, criteria. Criterion-referenced approaches are best suited for tests whose purpose is to determine mastery or absolute (as opposed to relative) levels of performance. Expert judgements from subject-matter experts (e.g., research scientists or policy specialists), experiential experts (e.g., fire fighters and other relevant practitioners), or representatives of key groups (e.g., female police officers, or aboriginal community members who are also forest fire fighters) may be used in setting the criteria. Common examples are the Angoff (and its various modifications), Nedelsky, and bookmark methods (Cizek, 2012; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006).

It should be noted that the term ‘experiential experts’ is not, to our knowledge, widely used in the assessment and standard setting literature but the concept is useful in both expanding the notion of ‘subject matter expert’ and also more fully articulating the varieties of expertise involved in such judgements. An expert takes information accumulated from their own (research) work and experience and combines it with the information offered by the descriptor of the test standard evidence and then arrives at an opinion as to the cut-score that should be drawn from all the assessment information provided to them. The work or experience that an expert draws from may be from having done, for example, research on the test domain such as applied physiologists’ research on physical fitness tasks or from having worked in that environment such as fire chiefs. The former would be described as a subject matter expert and the latter an experiential expert. The researcher conducting the standard setting study must then decide

whether to accept or reject, or what weight they wish to give to the expert's opinion as to the appropriate cut scores. Depending on the test purpose or type of task, standard setting researchers may give more evidence to experiential experts than subject matter experts. For example, experiential and subject-matter experts may differ in their opinion of a "minimal acceptable standard" based on how they perceive the rather subjective/conceptual "urgency" of the task. For example, dipping in to the resources of their experiences as a fire-fighter or fire chief, an experiential expert may be able to discern that a minimal acceptable standard for a task with little urgency will be different from one with a contextual scenario of "life-or-death". At the very least, however, distinguishing between the two types of experts clarifies their role and more clearly delineates the type of experience they are drawing upon for their recommendation. Likewise, the diversity of work and experience, and the interaction among these types of experts on the same standard setting panel, would be useful in informing where a cut-off should be set.

The bookmark approach is particularly well-suited for physical fitness testing (Rogers et al., 2014) because typical tests are comprised of a series of physical tasks and the total time to complete the test is recorded. The cut score is then set at the minimally acceptable level for, for example, 'safe, efficient and reliable' work performance and the panel members would be directed to, for example, look for the work rate that would be consistent with a minimally acceptable work capacity. A panel of experiential experts (e.g., fire fighter supervisors) is assembled. The panel members set the standard by reviewing prototypal test performances ('actors' are videotaped completing the circuit) and deciding on the level of performance (the completion time) that will be considered minimally acceptable. Because the prototypal test performances can be ordered in terms of test score, in this case total time to complete the circuit, a judge would place a bookmark at the point in the series of prototypal performances that they

consider the minimum acceptable performance. There are several ways for the judges to do this but all of them involve multiple rounds of judgments and the opportunity for panelists to re-consider the placement of their ‘bookmark’ in the ordered sequence of prototypical performances. For example, in the Rogers et al. standard setting a panel of 25 experienced Canadian Forces firefighter supervisors set the cut-score for the test. The panel members sat at individual computer stations within a large room with a DVD and worksheet were provided for each member to record the placement of their bookmark for up to three rounds.

The key aspect of choosing panel members is that they need to reflect the test purpose. A panel may be diverse and represent different viewpoints from the field in terms of supervisors, content experts, people dealing with liability, and people currently in the occupation. Diversity in panel membership is often considered a benefit. Panel members may be involved in describing the performance standard and/or setting the cut-score. Specific qualifications for these two tasks may be different. Experiential experts, as well as subject-matter experts, may be particularly useful as providing useful and rich verbal description of the performance standard as it is being developed. For example, the familiarity of the experiential experts with the target population of test-takers, and the use of the test, may be useful in keeping the performance standard description realistic. As Kane (1994) notes:

“... given the wide-ranging impact of the policy decisions involved in setting standards for high-stakes tests, it is probably important to have broad representation from groups with an interest in the stringency of the standard. However, this interest in having broad participation in the standard-setting process may be in conflict with the requirement that the judges be qualified to make the kind of decision they are being asked to make, and, therefore, a judicious trade-off may be called for” (p. 441).

Finally, the number of panel members, the number of tasks or performances they are rating, and the number of times they are rating them should be large enough to achieve an acceptably small

standard error of measurement for the resulting cut-score. Rogers et al. had 25 panel members who rated nine video-taped prototypical performances over three rounds.

It should be noted that standard setting methods based mainly on the judgements of experiential or subject matter experts is, in its essence, a type of psychological measurement (Nichols, et al. 2010) and psychophysical scaling that can be informed by contemporary cognitive science. The planning of panelists' activities need to take in to account recent findings in the cognitive psychology of judgment and decision-making; in particular the so-called fast and slow cognitive systems of the dual-process and dual-system theories in both cognitive and social psychology (Evans and Stanovich, 2013; Kahneman, 2011). The key message from contemporary cognitive science is that standard setting needs to recognize that panelists, like all humans, have two cognitive systems; one automatic and intuitive, the realm of systematic biases, the other conscious and deliberative. Kahneman's observations, in particular, remind us that panelists need more than one opportunity to engage their more conscious and deliberative thinking systems in re-assessing the cut-score they have selected in an earlier round. This deliberation is fundamental to the cognitive processing and decision making.

As Kane (1994) clearly states, the interpretation of cut-scores that are mainly based on subject-matter and/or experiential experts judgments about tasks or test performance depends on two assumptions. The first assumption is that the cut-score matches the specified performance standard, in the sense that test-takers with scores above the cut-score are likely to meet the performance standard and likewise that test-takers with scores below the cut-score are not likely to meet the performance standard. The second key assumption is that the specified performance standard is reasonable, given the purpose of the test. Test validation is then driven by evaluating the defensibility of these two assumptions in terms of the correspondence between the

procedures used to set the cut-scores and the purpose of the decision, the internal consistency of the results, and (very importantly) comparisons to external criteria. A recent standard setting exercise by Rogers et al. (2014) serves as a paradigmatic exemplar of best practices of judgement based standard setting.

The Importance of Conditional Estimates of Precision in Test Scores

As Zumbo and Rupp (2004) note, scoring test data eventually brings about consequences for test-takers and these consequences are mathematically dependent upon accurately estimating the error associated with test-takers' scores -- which is most crucial for test-takers with an observed score somewhere around the cut-score. It has long been recognized in the statistical theory of psychometrics that the score error is not constant along the continuum of test scores even though in early work in classical test theory unconditional raw-score standard error of measurement was reported and used. Evidence from different estimation methods has been accumulated in the last decade to support this. Generally speaking, for observed score models curves depicting the conditional standard error of measurement will be either: (a) somewhat inverse U-shaped, with smaller standard errors near the upper and lower tails of the score continuum and larger standard errors in the center of the score continuum, or (b) U-shaped with, with larger standard errors near the upper and lower tails of the score continuum and smaller standard errors in the center of the score continuum. Which one arises is likely due, in good part, to test design, test purpose, and test-taker population. Either way, best practices necessitate that conditional measures of precision need to be considered. So, rather than an overall standard error of measurement, one needs to estimate the standard error of measurement at relevant points on the test score distribution – such as at or near the cut-score. Moreover, it is clear that a

conditional raw-score standard error of measurement should be used for fair decision-making based on raw scores and that a conditional scale-score standard error of measurement should be reported if raw scores are transformed via linear or non-linear transformations to some other practically meaningful scale such as percentiles. In terms of quantifying or indexing reliability within an observed score context of conditional standard error, this has most notably resulted in a synthesis of conditional standard error estimation approaches for generalizability-theory designs (Brennan, 1998; 2001).

How Standards are tied to Validity Theory and a Focus on Test Consequences

Cronbach (1988), Kane (1992, 2006), Shepard (1993) and others advocate using argument as a way to frame or focus validation efforts and to clarify intended interpretations and uses. As Kane notes, the main advantage of the argument-based approach to validation is the guidance it provides in allocating research effort and in gauging progress in the validation effort (Kane, 2006, p. 23). He describes his view of validation as:

“To validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the proposed conclusions and decisions ... Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified.” (p. 17)

Kane conceptualizes validation as involving an interpretive argument and a validity argument. The interpretative argument is meant to provide a clear statement of the inferences and assumptions inherent in the proposed interpretations and uses of test results. These inferences and assumptions are to be evaluated in a series of analyses and empirical studies. As Kane states, an interpretive argument specifies the proposed interpretations and uses of scores by

laying out a chain or network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the scores. The validity argument provides an evaluation of the interpretive argument's coherence and of the plausibility of its inferences and assumptions. Kane states, "[w]hile the interpretations ... are evaluated in terms of their coherence and plausibility, decisions are evaluated in terms of their outcomes, or consequences. Figure 2 depicts the scoring, generalization, and extrapolation inferential chain linking, for example, observed test performance on a physical fitness test to the test use and decisions.

Insert Figure 2 about here

Test validation then focuses on a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard. Let us provide a brief example of how using an argument-based approach helps shine a light on the various inferences and assumptions. The inferences can also be informed by the multidimensional framework described by Zumbo (2007). Kane's (2013) interpretative/use argument for a physical fitness test for a job typically involves three main inferences. First, each test-taker's performance is recorded as time to complete the physical fitness circuit, yielding a 'test score'. The warrant for this scoring inference is the background research on the test tasks and the mediators and moderators of task performance. Second, the interpretation of the test score is then extended from a claim about expected performance over the test domain to a claim about the test's level of physical fitness as applied in the job. The extrapolation inference extends the score interpretation from a claim about the physical fitness as reflected in the test performance to a claim about the ability to apply

the level of fitness in practice. The warrant for the extrapolation inferences asserts that candidates who can effectively apply their physical fitness to complete the test tasks will generally also be able to effectively apply their physical fitness on a job, and even more importantly, that test-takers who cannot apply their physical fitness to complete the test tasks will generally also not be able to effectively apply their physical fitness on the job. The extrapolation inference does not change the value of the scaled score, but it does extend its interpretation from the testing context to the practice context. It is important to highlight that this is part of the argument makes a strong (empirically testable) assumption that the test score (i.e., the performance time) are not overly influenced by sources of construct irrelevant variance. Third, a decision is made about passing a test-taker by comparing the test-taker's test score to the cut score. This comparison and decision making is the same for all test-takers irrespective of any other test-taker characteristics. The warrant for this decision inference is usually a straightforward pass/fail decision rule. The validity argument provides an evaluation of this interpretive argument's coherence and the plausibility of its inferences and assumptions.

What is particularly noteworthy about the interpretative/use argument as applied to physical fitness testing for employment purposes described above is that it does not necessitate a predictive inference from a client's test score to a criterion measure of the test-taker's future performance in practice. As Hubley et al. (2014, p. 197) note, in contemporary practice there are many cases in which the evidence being presented should be treated as convergent evidence rather than criterion-related evidence. This puts a great deal of weight upon the evidence for the appropriateness of the cut-score. As we note above, this evidence comes in the form of statistical data from a normative sample (along with supplemental information) and/or a panel of subject-matter and/or experiential experts. As Kane (1994, 2001) and others (Hambleton and Pitoniak,

2006) repeatedly reminds us, this sort of judgmental panel evidence is likely to be convincing to the extent that: (a) the experts making the judgments are competent to make the necessary judgments, (b) the judgments are internally consistent across different tasks and/or across replication of the standard-setting process, and (c) judges are not subject to any identifiable source of bias.

Kane's model, however, does not explicitly deal with consequences of test use and decisions. Hubley and Zumbo (2011) introduced a new conceptualization of test validity that draws a clearer attention to the inherent elements (i.e., values), external influences (i.e., social consequences and side effects), and the process of validation in validity. The framework is meant to highlight the test's purpose and use, and its concomitant values, social and personal consequences, as well as social and personal side effects.

The Hubley-Zumbo framework, as adapted for physical fitness testing for employment purposes, is depicted in Figure 3. The model highlights several key features. First, at the core one can envision that from theories of test performance (i.e., the theories of the mediators and moderators of task performance arising from applied physiology) one develops tasks and a test, to which one ascribes test score inferences, interpretations, and use. From this test score inference and use emerges (a) intended social and personal consequences, and/or (b) unintended social and personal side effects. Very importantly, these consequences and/or side effects (either personal and/or social) may also influence test score inferences and use. Second, test score inference and use is effected and shaped by several forms of validity evidence, including but not necessarily limited to convergent/discriminant, known groups, job and task analysis, generalizability/invariance evidence, and most importantly standard setting and deriving the cut-score. Third, the dashed circle encompasses what we could consider construct validity –

containing within it the process of validation that provides the various sources of validity evidence and includes consequences and side effects. The centrality of the large (dashed) circle is meant to signify that construct validity is at the core of a unified view of validity and validation.

Insert Figure 3 about here

Following the Hubley-Zumbo model, what one validates are the inference, interpretation, action, or decision that one makes based on a test score and so validity is about the degree to which the inference is appropriate, meaningful, and useful given the individual or sample one is dealing with and the context in which one is working. 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, one aims to identify the degree to which construct underrepresentation and construct-irrelevant variance are problems. Because changes occur over time (e.g., in our empirical knowledge, theoretical understandings, values, society), the process of validation is an ongoing one. The ongoing nature is particularly relevant to standard setting and deriving a cut-score, which clearly not set once and for always but needs to be revisited as the key elements of the validity argument change, including societal values expectations for job performance.

Closing Remarks

Developments in standard setting methodology have largely gone on outside of the burgeoning field of physical fitness testing for job performance. The paper provided an overview of standard setting and deriving cut-scores, with an emphasis on contrasting and comparing methods, describing new methods and frameworks for best practices. The case is made, although at times subtly so, that in terms of best practices, establishing a test standard can be conceptualized as evidence/data-based policy making that is essentially tied to test validity. It should be noted that because of the voluminous measurement and assessment literature that has accumulated over the last 30 years, throughout the focus of this paper was on foundational issues and broad overview with an eye toward informing best practices.

Table 1 provides an overall summary of the steps in standard setting and deciding on a cut score. A key point highlighted in this paper is that one starts with the performance standard; which is the conceptual version of the desired level of competence. The passing score is the operational version of the desired level of competence. It is important to fully articulate the performance standard before you set the cut-score (the passing score). Setting a cut score in physical fitness testing can be done by either focusing on a test score distribution from a sample of test takers, which is supplemented with external information, relying on experiential or subject-matter expert panels, or some hybrid of these approaches. Validation studies begin next with considerations of misclassification of pass and failure decisions. Highlighting that setting a cut score is an empirically based policy decision; the second to last step is the recommendation to the policy maker. The process is wrapped up with the documentation of the process as well as the validation and consequences of the cut score. The evidential trail is key to the whole process because it is widely recognized that 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 (Kane,

1994). As Kane (1994, 2001) and others (Hambleton and Pitoniak, 2006) repeatedly reminds us, this sort of judgmental panel evidence is likely to be convincing to the extent that: the experts making the judgments are competent to make the necessary judgments, the judgments are internally consistent, consistent across replication of the standard-setting process, and judges are not subject to any identifiable source of bias.

Insert Table 1 about here

The field of physical fitness testing with its historic emergence from and emphasis on the applied physiology of job analysis, physiological investigation of mediators of task performance such as equipment, load, training, and coaching (rather than psychology’s focus on the statistical theory of test scores) developed a general approach to standard setting that can be described in the terms of measurement theory as a generic assessment procedure that aims to yield high precision and defensibility around the cut-score so that physical fitness tests are designed to fit the standard. Building standard setting in to test design is seen in several key physical fitness tests (see, for example, Jamnik, et al. 2010; Petersen, et al. 2013). A call for this sort of test development in standard setting dates back to over 20 years (e.g., Kane, 1994).

In closing, it is time to fundamentally rethink how we approach standard setting. Kane (1994) articulates this nicely and nudges us along when he addresses a key point that is largely ignored in discussions and the practices of standard setting. Kane states:

“So, we have a fairly high degree of ambiguity or arbitrariness in our choice of passing score. In most cases, it appears that there is no specific passing score that can be considered the "correct" passing score, and as a result we cannot demonstrate that we have chosen the correct passing score. The best that we can

do in supporting the choice of a performance standard and an associated passing score is to show that the passing score is consistent with the proposed performance standard and that this standard of performance represents a reasonable choice, given the overall goals of the assessment program. In practice, however, we seldom, if ever, achieve even this goal. A more modest, but realistic, goal in most cases is to assemble evidence showing that the passing score and its associated performance standard are not unreasonable.” (pp. 436-437).

Kane is suggesting that we stop thinking of setting a cut-score in absolute terms and as drawing a line in the sand such that the decision and its resulting consequences are permanently decided and irreversible. There is no perfect cut-score and only by chance is any test score the "true" score for a test taker, therefore there are real bounds on the reproducibility of the test score decisions (pass/fail) and on the cut-scores. In short, it is time to rethink the traditional absolute approach to cut-scores and oriented ourselves to a more realistic approach based on probability and the strength of an evidential trail to make a decision. As such, in terms of best practices, establishing a test standard and cut-score involves conceptualized it as evidence/data-based policy making that is essentially tied to test validity and establishing an evidential trail that supports that the proposed performance standards and cut-scores are not unreasonable and are reproducible and generalizable, akin to more widely accepted day-to-day scientific practices. There are very recent advances in this approach that involve adapting psychometric methods (e.g., generalizability theory) such as those described in Kannan et al. (2015).

References

- Brennan, R. L. 1998. Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, 22, 307-331.
- Brennan R. L. 2001. *Generalizability theory*. New York: Springer-Verlag.
- Cizek, G. J. 2012 (Ed.), *Setting performance standards: Foundations, methods, and innovations*. New York, NY: Routledge.
- Cizek, G. J., & Bunch, M. B. 2007. *Standard setting—A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Cronbach, L. J. 1988. Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
- Evans, J. St. B. T., & Stanovich, K. E. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8, 223–241.
- Glass, G. V. 1978. Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.
- Gledhill N., & Jamnik V.K. 2011. *Report on the development and validation of a Canadian physical performance exchange standard for type 1 wildland forest fire fighters*. Canadian Interagency Forest Fire Center, Inc
- Hambleton, R.K. 1984. Criterion-referenced measurement. In Husen & Postlethwaite (Eds.) *International Encyclopedia of Education*. Pergamon Press, NY.

Hambleton, R. K. 2001. Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*, (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. 2006. Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Washington, DC: American Council on Education.

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. 2014. Synthesis of Validation Practices in Two Assessment Journals: Psychological Assessment and the European Journal of Psychological Assessment. In Bruno D. Zumbo, and Eric K.H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 93–213). New York: Springer.

Hubley, A. M., & Zumbo, B.D. 2011. Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement*, 103, 219–230.

Jamnik, V. K., Gumienak, R., & Gledhill, N. 2013. Physiological Employment Standards II: Developing and Implementing Physical Employment Standards for Safety-Related Occupations. *European Journal of Applied Physiology*, 113, 2447–2457.

Jamnik, V. K., Thomas, S. G., Burr, J. F., & Gledhill, N. 2010. Construction, Validation and Derivation of Performance Standards for a Fitness Test for Correctional Officer Applicants. *Applied Physiology, Nutrition and Metabolism*, 35, 59–70.

Kahneman, D. 2011. *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kane, M. 1992. An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535.

Kane, M. 1994. Validating the Performance Standards Associated With Passing Scores. *Review of Educational Research*, 64, 425-461.

Kane, M. 1998. Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, 5, 129-145.

Kane, M. 2001. "So Much Remains the Same: Conception and Status of Validation in Setting Standards." In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*, (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. 2006. Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport: American Council on Education/Praeger.

Kane, M. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73.

Kannan, P., Sgammato, A., Tannenbaum, R. J., & Katz, I.R. 2015. Evaluating the Consistency of Angoff-Based Cut Scores Using Subsets of Items Within a Generalizability Theory Framework. *Applied Measurement in Education*, 28, 169-186.

Livingston, S. A., & Zieky, M. J. 1982. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

McKinley, D.W., Boulet, J. R., & Hambleton, R. K. 2005. A Work-centered approach for setting passing scores on performance-based assessments. *Evaluation & the Health Professions*, 28, 349-369

Messick, S. 1989. Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Nichols, P., Twing, J. Pearson, Mueller, C. D., & O'Malley, K. 2010. Standard-Setting Methods as Measurement Processes. *Educational Measurement: Issues and Practice*, 29, 14–24.

Petersen, S. R., Docherty, D., & Rogers, W. T. 2013. *Revision of the Standard for the Canadian Forces Firefighter Physical Fitness Maintenance Evaluation (FF PFME)*. Report Submitted to the Department of National Defence. Edmonton: University of Alberta.

Rogers, W.T., Docherty, D., & Petersen, S. R. 2014. Establishment of performance standards and a cut-score for the Canadian Forces Firefighter Physical Fitness Maintenance Evaluation (FF PFME). *Ergonomics*, 57, 1750-1759,

Rogers, W. T., & Ricker, K. L. 2006. Establishing Performance Standards and Setting Cut-scores. *Alberta Journal of Education*, 52, 16–24.

Shepard, L. 1993. Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Tipton, M. J., Milligan, G. S., & Reilly, T. J. 2013. Physiological employment standards I. Occupational fitness standards: objectively subjective?. *European Journal of Applied Physiology*, 113, 2435-2446.

Zieky M. J. 2001. So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B.D. 2001. Methodology and Measurement Matters In Establishing A Bona Fide Occupational Requirement For Physically Demanding Occupations. In Norman Gledhill, Jean

Bonneau, & Art Salmon (Eds.). *Proceedings of the Consensus Forum on Establishing BONA FIDE Requirements for Physically Demanding Occupations*, (pp. 37-52). Toronto, ON.

Zumbo, B. D. 2007. Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Psychometrics*, (Vol. 26, pp. 45–79). Amsterdam: Elsevier.

Zumbo, B. D. 2009. Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. 2002. The Construction And Use Of Psychological Tests And Measures. In the Psychology theme of the *Encyclopedia of Life Support Systems (EOLSS)*, Eolss Publishers, Oxford, UK.

Zumbo, B. D., & Rupp, A. A. 2004. Responsible Modeling of Measurement Data For Appropriate Inferences: Important Advances in Reliability and Validity Theory. In David Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Table 1. Summary of steps in standard setting and deciding on a cut score.

Step 1: Prepare a description of the standard; the performance level descriptor

Step 2: Choose an appropriate method for setting the cut-score based on the nature and purpose of the test.

Norm-referenced test interpretation

- Obtain a sample of test takers.
- Supplement data from the test score distribution with external criteria such as the descriptor of the performance levels in step 1, the results of a job analysis, and/or a reference group who take the test but are already performing the job.
- Obtain a recommended cut score.

Criterion-based test interpretation

- Choose a panel of experiential and subject-matter experts.
- Train panelists about the standard described in step 1. Training should include information about exemplar or borderline test takers.
- Collect ratings from panelists.
- Conduct discussion among panelists and provide feedback.
- Collect, at least, another two rounds or ratings from panelists, including discussion and feedback after each round.
- Obtain a recommended cut score.
- Collect panelist feedback.

Step 3. For either class of methods, one needs to consider misclassifications, and the balance of the trade-off of incorrect pass versus incorrect failure.

Step 4: Make recommendations to a policy maker about the pass score.

Step 5: Document all of the steps and information in steps 1 and 2; and begin the phase of validation of the cut-score and consideration of consequences of decision making.

Figure 1. A Description of Decision Theoretic Outcomes

Figure 2. Kane's Argument-based Approach to Validation

Figure 3. Hubley and Zumbo's Model of Validity and Validation Adapted for Physical Fitness

Testing for Employment. Modified from Hubley and Zumbo (2011); used with permission.

Draft

		<u>Test Decision</u>		
		Fail	Pass	
Decision based on a Gold Standard	Fail	True Negatives	False Positives	Specificity
	Pass	False Negatives	True Positives	Sensitivity
		Negative Predictive Value	Positive Predictive Value	

Specificity = true negatives / (true negatives + false positives)

Sensitivity = true positives / (false negatives + true positives)

Negative Predictive Value (NPV) = true negatives / (true negatives + false negatives)

Positive Predictive Value (PPV) = true positives / (false positives + true positives)