ГОСУДАРСТВЕННЫЙ ЦЕНТРАЛЬНЫЙ ОРДЕНА ЛЕНИНА
ИНСТИТУТ ФИЗИЧЕСКОЙ КУЛЬТУРЫ

Доктор педагогических наук,
профессор В. М. ЗАЦИОРСКИЙ

СПОРТИВНАЯ МЕТРОЛОГИЯ

ПЕДАГОГИЧЕСКИЙ КОНТРОЛЬ В ТРЕНИРОВОЧНОМ ПРОЦЕССЕ

(основы теории тестов и оценок)

Учебное пособие для студентов
институтов физической культуры

Москва 1978

ГОСУДАРСТВЕННЫЙ ЦЕНТРАЛЬНЫЙ ОРДЕНА ЛЕНИНА
ИНСТИТУТ ФИЗИЧЕСКОЙ КУЛЬТУРЫ

Доктор педагогических наук,
профессор В. М. ЗАЦИОРСКИЙ

СПОРТИВНАЯ МЕТРОЛОГИЯ

ПЕДАГОГИЧЕСКИЙ КОНТРОЛЬ В ТРЕНИРОВОЧНОМ ПРОЦЕССЕ

(основы теории тестов и оценок)

Учебное пособие для студентов
институтов физической культуры

Москва 1978

Настоящее учебное пособие включает два раздела: 1. Основы теории тестов и 2. Основы теории педагогических оценок.

В последующих выпусках будут представлены материалы по другим проблемам педагогического контроля в спорте.

ОСНОВЫ ТЕОРИИ ТЕСТОВ

Основные понятия

Тестом называется измерение или испытание, проводимое на спортсмене с целью определения его состояния. Процесс испытаний называется тестированием, полученное в итоге измерения числовое значение — результатом тестирования (или результатом теста). Например, бег 100 м — это тест, процедура проведения забегов и хронометража — тестирование, время бега — результат теста.

Тесты, в основе которых лежат двигательные задания, называют двигательными (или моторными) тестами. В этих тестах в качестве результатов могут выступать либо двигательные достижения (время прохождения дистанции, количество повторений, преодоленное расстояние и т. п.), либо физиологические и биохимические показатели. В зависимости от этого, а также от задания, которое стоит перед испытуемым, различают три группы двигательных тестов (табл. 1).

В тех случаях, когда используется не один, а несколько тестов, имеющих единую конечную цель (например, оценку состояния спортсмена в соревновательном периоде тренировки), такая группа называется комплексом или батареей тестов.

Не всякие измерения могут быть использованы как тесты, для этого необходимо выдержать специальные требования. К ним относятся:

1. Стандартность — процедура и условия тестирования должны быть одинаковыми во всех случаях применения теста.
2. Наличие системы оценок (раздел «Основы теории педагогических оценок»).
3. Надежность теста.
4. Информативность теста.

Тесты, удовлетворяющие требованиям надежности и информативности, называют добротными.

Таблица 1.

Разновидности двигательных тестов

Название	Двигательное достижение, которое должен показать испытуемый	Результат теста	Пример
1. Контрольное упражнение	Максимальное	Двигательное достижение	Бег 1500 м, время бега
2. Функциональные пробы	Одинаковое для всех	Физиологические или биохимические показатели	Регистрация потребления O_2 при стандартной работе, 1000 кгм/мин
3. Максимальные функциональные тесты	Максимальное	Физиологические или биохимические показатели	Определение максимального кислородного долга или максимального потребления кислорода

Надежность тестов

Понятие о надежности тестов. Надежностью тестов называют степень совпадения результатов при повторном тестировании одних и тех же людей (или других объектов) в одинаковых условиях. В идеале один и тот же тест, примененный к тем же испытуемым в тех же самых условиях, должен давать одинаковые результаты (если состояние испытуемых не изменилось). Однако даже при самой строгой стандартизации испытаний и точной аппаратуре результаты тестирования изменяются от попытки к попытке. Например, спортсмен, только что выжавший на кистевом динамометре 55 кг, через несколько минут покажет лишь 50 кг. Результаты тестирования всегда не-

сколько варьируют. Подобную вариацию называют внутрииндивидуальной или, используя более общую терминологию математической статистики, внутриклассовой. Четыре основные причины вызывают эту вариацию:

1. Изменение состояния испытуемых (утомление, врабатывание, научение, изменение мотивации, концентрация внимания и т. п.).

2. Неконтролируемые изменения внешних условий и аппаратуры (температура, ветер, влажность, изменения напряжения в электросети, присутствие посторонних лиц и т. п.).

3. Изменение состояния лица, проводящего или оценивающего тест (и, конечно, замена одного экспериментатора или судьи другим).

4. Несовершенство теста (есть такие тесты, которые заведомо мало надежны, например, штрафные броски в баскетбольную корзину до первого промаха. Даже баскетболист, имеющий высокий процент попадания, может случайно ошибиться на первых бросках).

Чтобы разобраться в идее методов, используемых для суждения о надежности тестов, рассмотрим упрощенный пример. Предположим, что мы хотим сравнить результаты прыжков в длину с места у двух спортсменов по двум выполненным попыткам. Мы хотим сделать точные выводы, поэтому не ограничиваемся регистрацией лишь лучших результатов. Допустим, что результаты каждого из спортсменов варьируют в пределах ±10 см от средней величины и равны соответственно 220+10 см (т. е. 210 и 230 см) и 320+10 см (т. е. 310 и 330 см). Если мы встретимся с таким случаем, то вывод, конечно, будет совершенно однозначным: второй спортсмен превосходит первого. Различия между спортсменами (320 см — 220 см = =100 см) явно больше случайных колебаний результатов (±10 см). Гораздо менее определенным будет вывод, если при той же самой внутриклассовой вариации (равной ±10 см) различие между испытуемыми (межклассовая вариация) будет маленьким. Скажем, средние значения будут равны 220 см (в одной попытке 210 см, в другой 230 см) и 222 (212 и 232 см). Тогда может случиться, например, что в первой попытке первый спортсмен прыгнет 230 см, а второй — только 212, и создастся впечатление, что первый существенно сильнее второго.

Из примера видно, что основное значение имеет не са-

ма по себе внутриклассовая изменчивость, а ее соотношение с межклассовыми различиями. Одна и та же внутриклассовая вариация дает разную надежность при разных различиях между классами (в частном случае испытуемыми).

Теория надежности теста исходит из того, что результат любого измерения (x_t), проводимого на человеке, есть сумма двух величин:

$$x_t = x_\infty + x_e \qquad (1),$$

где x_∞ — это так называемый истинный результат, который мы хотим зафиксировать;

x_e — ошибка, вызванная неконтролируемой вариацией в состоянии испытуемого, привносимая измерительным прибором и др.

Под истинным результатом по определению понимают среднее значение x_t при бесконечно большом числе наблюдений в одинаковых условиях (поэтому при X и ставят знак бесконечности ∞).

Если ошибки случайны (их сумма равна нулю и в разных попытках они не зависят друг от друга), тогда из математической статистики следует:

$$\sigma^2_t = \sigma^2_\infty + \sigma^2_e \qquad (2),$$

т. е. зарегистрированная в опыте дисперсия результатов (σ^2_t) равна сумме дисперсий истинных результатов (σ^2_∞) и ошибок (σ^2_e).

σ^2_∞ характеризует идеализированную (т. е. свобод-

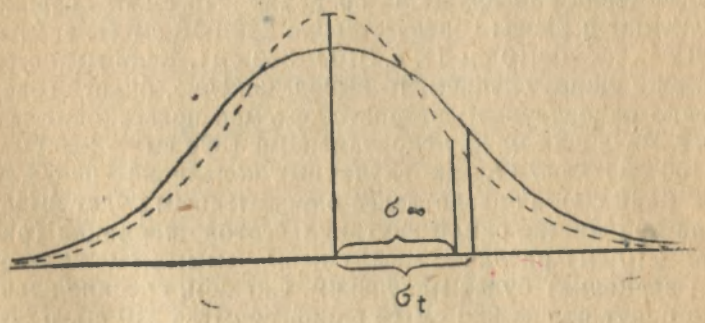


Рис. 1. Распределение зарегистрированных результатов теста (x_t, сплошная линия) и теоретических истинных результатов (x_∞, пунктир)

Средние величины преполагаются равными, $\sigma_t > \sigma_\infty$

ную от ошибок) межклассовую вариацию, σ^2_e — внутриклассовую изменчивость. Влияние σ^2_e изменяет распределение результатов теста (рис. 1).

По определению коэффициент надежности (r_{tt}) равен отношению истинной дисперсии к дисперсии, зарегистрированной в опыте:

$$r_{tt} = \frac{\text{истинная дисперсия}}{\text{зарегистрированная дисперсия}}$$

$$r_{tt} = \frac{\sigma^2_\infty}{\sigma^2_t} = \frac{\sigma^2_t - \sigma^2_e}{\sigma^2_t} = 1 - \frac{\sigma^2_e}{\sigma^2_t} \qquad (3).$$

Иными словами, r_{tt} есть просто доля истинной вариации в той вариации, которая зарегистрирована в опыте.

Кроме коэффициента надежности, используют еще индекс надежности:

$$r_{t\infty} = \sqrt{r_{tt}} \qquad (4),$$

который рассматривают как теоретический коэффициент корреляции зарегистрированных значений теста с истинными.

Оценка коэффициента надежности по экспериментальным данным. Понятие об истинном результате теста является абстракцией.

В опыте σ^2_∞ измерить нельзя (ведь нельзя же в действительности провести бесконечно большое число наблюдений в одинаковых условиях). Поэтому приходится использовать косвенные методы.

Наиболее предпочтителен для оценки надежности дисперсионный анализ с последующим расчетом так называемых внутриклассовых коэффициентов корреляции.

Дисперсионный анализ, как известно, позволяет разложить зарегистрированную в опыте вариацию результатов теста на составляющие, вызванные влиянием отдельных факторов. Например, если зарегистрировать у испытуемых их результаты в каком-либо тесте, повторяя этот тест в разные дни, причем в каждый из дней делать по нескольку попыток, периодически меняя экспериментаторов, то будет иметь место вариация:

а) от испытуемого к испытуемому (межиндивидуальная вариация),

б) от дня ко дню,

в) от экспериментатора к экспериментатору,

г) от попытки к попытке.

Дисперсионный анализ дает возможность выделить и оценить вариации, вызванные этими факторами.

Рассмотрим на упрощенном примере, как это делается. Предположим, что мы измерили у 5 испытуемых результаты двух попыток ($k=5$, $n=2$):

Испытуемые	Попытка	
	первая	вторая
1	10	6
2	8	9
3	7	5
4	4	7
5	2	3

Результаты дисперсионного анализа (см. курс математической статистики) приведены в традиционной форме (табл. 2).

Таблица 2

Результаты дисперсионного анализа

Вариация	Сумма квадратов	Степени свободы	Дисперсия
Межклассовая (между испытуемыми)	$Q_1=45{,}4$	$f_1=(k-1)=4$	$\sigma^2_1=Q_1:f_1=11{,}35$
Внутриклассовая (внутрииндивидуальная, остаточная)	$Q_2=15{,}5$	$f_2=nk-k=5$	$\sigma^2_2=Q_2:f_2=3{,}1$
Общая	$Q=60{,}9$	$f=nk-1=9$	$\sigma^2=Q:f=6{,}77$

Надежность оценивается с помощью так называемого внутриклассового коэффициента корреляции:

$$r'_{tt}=\frac{\sigma^2_1-\sigma^2_2}{\sigma^2_1+\left(\frac{n}{n'}-1\right)\sigma^2_2} \qquad (5),$$

где r' — коэффициент внутриклассовой корреляции (коэффициент надежности), его обозначают с дополнительным штрихом — r', чтобы отличить от обычного коэффициента корреляции — r; n — использованное в тесте число попыток; n' — число попыток, для которого проводится оценка надежности.

Например, если мы хотим оценить по данным нашего примера надежность средней из двух попыток, то:

$$r'_{tt}=\frac{11{,}35-3{,}1}{11{,}35+\left(\frac{2}{2}-1\right)3{,}1}=0{,}727$$

Если ограничиться только одной попыткой, то надежность будет равна:

$$r_{tt} = \frac{11{,}35 - 3{,}1}{11{,}35 + \left(\frac{2}{1} - 1\right) 3{,}1} = 0{,}571;$$

а если мы увеличим число попыток до четырех, коэффициент надежности также несколько возрастет:

$$r'_{tt} = \frac{11{,}35 - 3{,}1}{11{,}35 + \left(\frac{2}{4} - 1\right) 3{,}1} = 0{,}764$$

Таким образом, чтобы оценить надежность, надо:

1) выполнить дисперсионный анализ;

2) рассчитать внутриклассовый коэффициент корреляции (коэффициент надежности).

Некоторые сложности возникают, когда имеет место так называемый тренд, т. е. систематическое повышение или понижение результатов от попытки к попытке (рис. 2). В этом случае используют более сложные методы оценки надежности (в настоящей работе они не описаны).

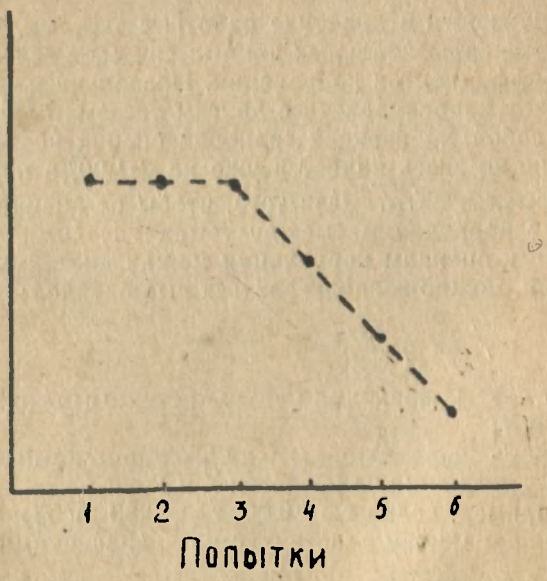


Рис. 2. Серия из шести попыток, из которых три последние подвержены тренду.

Для случая двух попыток и отсутствия тренда величины внутриклассового коэффициента корреляции практически совпадают со значениями обычного коэффициента корреляции между результатами первой и второй попытки. Поэтому в таких ситуациях для оценки надежности может использоваться и обычный коэффициент корреляции (он при этом оценивает надежность одной, а не двух попыток). Однако, если число повторных попыток в тесте больше двух и в особенности если используются сложные схемы тестирования (например, по 2 попытки в день в течение двух дней), только расчет внутриклассового коэффициента является правильным.

Коэффициент надежности не является абсолютным показателем, характеризующим тест. Этот коэффициент может изменяться в зависимости от контингента испытуемых (например, быть различным у начинающих и квалифицированных спортсменов), условий тестирования (проводятся ли повторные попытки одна за другой или, скажем, с интервалом в одну неделю) и других причин. Поэтому всегда надо описывать, как и на ком проводился тест.

Надежность в практике работы с тестами. Ненадежность экспериментальных данных снижает величину оценок коэффициентов корреляции. Поскольку ни один тест не может коррелировать с другим тестом больше, чем сам с собой, то верхней границей оценки коэффициента корреляции здесь является уже не ± 1.00, а индекс надежности $r_{t\infty}=\sqrt{r_{tt}}$. Для того, чтобы от оценки коэффициентов корреляции между эмпирическими данными перейти к оценкам корреляции между «истинными» значениями, можно воспользоваться выражением:

$$r_{xy}=\frac{r_{xy}}{\sqrt{r_{xx}+r_{yy}}} \qquad (6),$$

где r_{xy} — корреляция между «истинными» значениями X и У;

r_{xy} — корреляция между эмпирическими данными;

r_{xx} и r_{yy} — оценка надежности X и У.

Например, если $r_{xy}=0{,}60$, $r_{xx}=0{,}80$ и $r_{yy}=0{,}90$, то корреляция между «истинными» значениями равна 0,707.

Приведенная формула (6) называется коррекцией на уменьшение, она постоянно используется в практике.

Нет фиксированного значения надежности, после которого тест можно считать приемлемым. Все зависит от важности выводов, которые делаются на основе применения теста. Все же в большинстве случаев в спорте можно использовать следующие примерные ориентиры: 0,95—0,99 — отличная надежность; 0,90—0,94 — хорошая; 0,80—0,89 — приемлемая; 0,70—0,79 — плохая; 0,60—0,69 — для индивидуальных оценок сомнительная, тест пригоден лишь для характеристики группы испытуемых (а не отдельных спортсменов).

Добиться некоторого повышения надежности теста можно, увеличивая число повторных попыток. Вот как, например, возрастала надежность теста (метание гранаты 350 г с разбега) по мере увеличения числа попыток: 1 попытка — 0,53; 2 попытки — 0,72; 3 попытки — 0,78; 4 попытки — 0,80; 5 попыток — 0,82; 6 попыток — 0,84. Из примера видно, что если сначала надежность возрастает быстро, то после 3—4 попыток прирост существенно замедляется.

При нескольких повторных попытках можно определять результаты разными способами: а) по лучшей попытке, б) по средней арифметической, в) по медиане, г) по средней из двух или трех лучших попыток и т. п. Исследования показали, что в большинстве случаев наиболее надежным является использование средней арифметической величины, несколько менее надежна медиана, еще менее надежен лучший результат. Принятая в спорте практика определения победителя по результату лучшей попытки не является наиболее надежной. В педагогическом контроле рекомендуется использовать среднюю величину или медиану нескольких попыток.

Говоря о надежности тестов, различают:

1) стабильность (воспроизводимость), 2) согласованность, 3) эквивалентность тестов.

Стабильность теста. Под стабильностью понимают воспроизводимость результатов теста при его повторении через определенное время в одинаковых условиях. Повторное тестирование обычно называют ретестом. Схема оценки стабильности такова:

$$\text{тест} \xrightarrow[\text{интервал}]{\text{временный}} \text{ретест}$$

При этом различают два случая. В первом ретест прово-

дят для того, чтобы получить надежные данные о состоянии испытуемого в течение всего временного интервала между тестом и ретестом (например, чтобы получить надежные данные о функциональных возможностях лыжников в июне, у них проводят измерение МПК дважды с интервалом в одну неделю). В этом случае важны точные результаты теста и оценка надежности должна проводиться, как это описано выше, с помощью дисперсионного анализа.

В другом варианте может быть важным лишь сохранение порядка испытуемых в группе (остается ли первый первым, последний — среди последних). В этом случае стабильность оценивают по коэффициенту корреляции между тестом и ретестом.

Стабильность зависит от: 1) вида теста, 2) контингента испытуемых, 3) временного интервала между тестом и ретестом.

Например, морфологические характеристики при небольших временных интервалах весьма стабильны, наименьшую стабильность имеют тесты на точность движений (например, броски в цель).

У взрослых результаты тестирования более стабильны, чем у детей; у спортсменов стабильность выше, чем у не занимающихся спортом.

С увеличением временного интервала между тестом и ретестом стабильность снижается (табл. 3).

Таблица 3

Стабильность теста при разных временных интервалах (120 испытуемых студентов, в один день выполнялось по две попытки)

Тест	Ретест сразу по окончании теста	Ретест через 1 месяц
Бег 100 м	0,94	0,76
Прыжок в длину с места	0,93	0,82

Согласованность теста характеризуется независимостью результатов тестирования от личных качеств лица, проводящего или оценивающего тест*.

* Вместо термина «согласованность» довольно часто используют термин «объективность». Такое словоупотребление неудачно, так как совпадение результатов разных экспериментаторов или судей (экспертов) вовсе не говорит об их объективности. Они могут все вместе сознательно или несознательно ошибаться, искажая объективную истину.

Согласованность определяется по степени совпадения результатов, получаемых на одних и тех же испытуемых разными экспериментаторами, судьями, экспертами. Здесь есть два варианта:

1. Лицо (судья, эксперт) только оценивает результаты теста, не влияя на его выполнение. Например, одну и ту же письменную работу разные экзаменаторы могут оценить по-разному. Различаются оценки судей в гимнастике, фигурном катании, боксе, показатели ручного хронометрирования, оценка электрокардиограммы или рентгенограммы разными врачами и т. п.

2. Лицо, проводящее тест, влияет на результаты. Например, некоторые экспериментаторы более настойчивы и требовательны, чем другие, лучше мотивируют испытуемых. Это сказывается на результатах (которые сами по себе могут измеряться вполне объективно).

Согласованность теста — это по существу надежность оценки его результатов или проведения разными людьми.

Эквивалентность тестов. Нередко тест представляет собой результат выбора из определенного числа однотипных тестов. Например, броски в баскетбольную корзину можно выполнять с разных точек, спринтерский бег может быть на дистанции, скажем, 50, 60 или 100 м, подтягивания можно выполнять на кольцах или перекладине, хватом сверху или снизу и т. д. В таких случаях может использоваться так называемый метод параллельных форм, когда испытуемым предлагают выполнить две разновидности одного и того же теста и затем оценивают степень совпадения результатов. Схема тестирования здесь следующая:

$$\text{форма A} \xrightarrow[\text{временной интервал}]{\text{минимальный}} \text{форма B}$$

Рассчитанный между результатами тестирования коэффициент корреляции называют коэффициентом эквивалентности. Отношение к эквивалентности тестов зависит от конкретной ситуации. С одной стороны, если два теста или больше эквивалентны, их совместное применение повышает надежность оценок. С другой — может оказаться полезным оставить в батарее только один эквивалентный тест — это упростит тестирование и лишь незначительно снизит информативность комплекса тестов. Решение этого вопроса зависит от таких причин, как

сложность и громоздкость тестов, степень необходимой точности тестирования и т. п.

Если все тесты, входящие в какой-либо комплекс тестов, высоко эквивалентны, такой комплекс называется гомогенным. Весь этот комплекс меряет одно какое-то свойство моторики человека. Скажем, комплекс, состоящий из прыжков с места, в длину, вверх и тройного, вероятно, будет гомогенным. Наоборот, если в комплексе нет эквивалентных тестов, то все тесты, входящие в него, меряют разные свойства. Такой комплекс называется гетерогенным. Пример гетерогенной батареи тестов: подтягивания на перекладине, наклон вперед (для проверки гибкости), бег 1500 м.

Пути повышения надежности тестов. Надежность тестов до определенной степени может быть повышена путем:

а) более строгой стандартизации тестирования;

б) увеличения количества попыток;

в) увеличения числа оценщиков (судей, экспертов) и повышения согласованности их мнений;

г) увеличения количества эквивалентных тестов;

д) лучшей мотивации испытуемых.

Информативность тестов

Основные понятия. Начнем с примера. Допустим, что мы хотим определить уровень специальной силовой подготовленности спринтеров—бегунов и пловцов. Для этого мы предполагаем использовать такие показатели: 1) кистевая динамометрия, 2) сила подошвенного сгибания стопы, 3) сила разгибателей плечевого сустава (эти мышцы несут большую нагрузку при плавании кролем), 4) сила мышц—разгибателей шеи. Мы хотим на основе этих тестов управлять тренировочным процессом, в частности, находить слабые звенья двигательного аппарата и их целенаправленно укреплять. Хорошие ли тесты мы выбрали? Информативны ли они? Даже не проводя специальных экспериментов, можно догадаться, что тест № 2, вероятно, информативен у спринтеров-бегунов, тест № 3—у пловцов, а тесты № 1 и 4, наверно, ни у пловцов, ни у бегунов не покажут ничего интересного (хотя могут оказаться очень полезными в других видах спорта, например в борьбе). В разных случаях одни и те же тесты могут иметь разную информативность.

Информативность теста — это степень точности, с которой он измеряет свойство (качество, способность, характеристику и т. п.), для оценки которого используется. Информативность нередко называют также валидностью (от английского validity — обоснованность, действительность, законность).

Вопрос об информативности теста распадается на два частных вопроса: что измеряет данный тест? Как точно он это делает?

Например, можно ли по такому показателю, как максимальное потребление кислорода (МПК), судить о подготовленности бегунов-стайеров, и если можно, то с какой степенью точности? Иными словами, какова информативность МПК у стайеров?

Если тест используется для определения (диагноза) состояния спортсмена в момент обследования, то говорят о диагностической информативности. Если же на основе результатов тестирования хотят сделать вывод о возможных будущих показателях спортсмена, тест должен обладать прогностической информативностью. Тест может быть диагностически информативен, а прогностически нет, и наоборот.

Степень информативности может характеризоваться количественно на основе опытных данных (так называемая эмпирическая информативность) и качественно на основе содержательного анализа ситуации (содержательная или логическая информативность).

Эмпирическая информативность (случай первый — существует измеряемый критерий). Идея определения эмпирической информативности состоит в том, что результаты теста сравнивают с некоторым критерием. Для этого рассчитывают коэффициент корреляции между критерием и тестом (такой коэффициент называют коэффициентом информативности и обозначают r_{tk}. t первая буква в слове «тест», k — в слове «критерий).

В качестве критерия берется показатель, заведомо и бесспорно отражающий то свойство, которое мы собираемся мерить с помощью теста.

Нередко бывает так, что существует вполне определенный критерий, с которым можно сравнить предполагаемый тест. Например, при оценке специальной подготовленности спортсменов в видах спорта с объективно измеряемыми результатами таким критерием служит

обычно сам результат: тот тест более информативен, корреляция которого со спортивным результатом выше. В случае определения прогностической информативности критерием является показатель, прогноз которого надо осуществить (например, если прогнозируется длина тела ребенка, то критерием является длина его тела во взрослые годы).

Чаще всего в спортивной метрологии критериями служат:

1. Спортивный результат.

2. Какая-либо количественная характеристика основного спортивного упражнения (например, длина шага в беге, сила отталкивания в прыжках, успешность борьбы под щитом в баскетболе, выполнение подачи в теннисе и волейболе, процент точных длинных передач в футболе и т. д. и т. п.).

3. Результаты другого теста, информативность которого доказана (это делают, если проведение теста-критерия громоздко и сложно и можно подобрать другой тест столь же информативный, но более простой. Например, вместо газообмена определять просто частоту пульса). Этот частный случай, когда критерием является другой тест, называют конкурентной информативностью.

4. Принадлежность к определенной группе. Например, можно сравнивать членов сборной команды страны, мастеров спорта и перворазрядников; принадлежность к одной из этих групп является критерием. В данном случае используются специальные разновидности корреляционного анализа.

5. Так называемый составной критерий, например, сумма очков в многоборье. При этом виды многоборья и таблицы очков могут быть как общепринятыми, так и заново составлены экспериментатором (о том, как составляются такие таблицы, см. ниже). К составному критерию прибегают, когда нет единичного критерия (например, если стоит задача оценить общую физическую подготовленность, мастерство игрока в спортивных играх и т. п. — во всех этих случаях ни один показатель, взятый сам по себе, не может служить критерием).

Пример определения информативности одного и того же теста (скорость бега 30 м с хода у мужчин) при разных критериях приведен в табл. 4.

Таблица 4

Информативность теста «бег 30 м с хода»

Критерий	Мера критерия	Коэффициент информативности
1. Прыжок в длину с разбега	Результат прыжка, см	0,658
2. Разбег в прыжках в длину	Скорость бега на последних 10 метрах, м/с	0,918
3. Спортивные достижения в прыжках в длину	Разряд по легкой атлетике от второго до мастера спорта	0,715
4. Результат троеборья: бег 100 м, прыжки в длину, бег 110 м с/б	Сумма очков	0,764

Вопрос о выборе критерия является по существу самым важным при определении реального значения и информативности теста. Например, если стоит задача определить информативность такого теста, как прыжок в длину с места у спринтеров, то можно выбрать разные критерии: результат в беге на 100 м, длину шага, отношение длины шага к длине ног или к росту и т. п. Информативность теста при этом будет меняться (в приведенном примере она возрастала от 0,558 для скорости бега до 0,781 для отношения «длина шага /длина ноги»).

В видах спорта, где нельзя объективно измерить спортивное мастерство, стараются обойти эту трудность введением искусственных критериев. Например, в командных спортивных играх эксперты располагают всех игроков по их игровому мастерству в определенном порядке (т. е. составляют списки 20, 50 или, скажем, 100 сильнейших игроков). Место, занятое спортсменом (как говорят, его ранг), рассматривается в качестве критерия, с которым и сравнивают результаты тестов с целью определения их информативности.

Встает вопрос: зачем использовать тесты, если известен критерий? Например: не проще ли устроить контрольные соревнования и определить спортивный результат, чем определять достижения в контрольных упражне-

ниях? Применение тестов имеет следующие преимущества:

1) спортивный результат не всегда можно или целесообразно определить (например, нельзя часто проводить соревнования в марафонском беге, зимой обычно нельзя определить результат в метании копья, летом нельзя зарегистрировать результат в лыжных гонках и т. п.);

2) спортивный результат зависит от многих причин (факторов), таких, как сила спортсмена, его выносливость, техника и т. д. Применение тестов дает возможность определить сильные и слабые стороны спортсмена, оценить каждый из этих факторов в отдельности.

Эмпирическая информативность (случай второй — единичного критерия нет; факторная информативность). Нередко встречаются случаи, когда нет единичного критерия, с которым можно сравнить результаты предполагаемых тестов. Допустим, мы хотим найти наиболее информативные тесты для оценки силовой подготовленности молодежи. Что предпочесть: подтягивания на перекладине или отжимания в упоре на брусьях? Или, быть может, приседания со штангой, тягу штанги, либо подъемы в сед из положения лежа на спине? Что здесь может быть критерием правильного выбора теста?

Можно рассуждать так: предложим испытуемым большую батарею разнообразных силовых тестов, а затем отберем среди них те, которые дают наибольшую корреляцию с результатами всего комплекса (ведь нельзя же систематически пользоваться всем комплексом, он слишком громоздок и неудобен для этого). Эти тесты будут наиболее информативны: применяя их, мы получим сведения о возможных результатах испытуемых по всему исходному комплексу тестов. Но результаты в комплексе тестов не выражаются одним числом. Можно образовать, конечно, какой-либо составной критерий (например, определить сумму набранных очков по какой-либо шкале). Однако гораздо более эффективен другой путь, основанный на идеях факторного анализа.

Факторный анализ — один из методов многомерной статистики (слово «многомерный» указывает, что изучается одновременно много разных показателей, например, результаты испытуемых во многих тестах). Факторный анализ — довольно сложный метод и в институтах физической культуры не изучается (хотя в спортивной науке

используется очень широко). Мы ограничимся изложением лишь основной идеи этого метода.

Факторный анализ исходит из того, что результат любого теста является следствием одновременного действия ряда непосредственно ненаблюдаемых факторов. Например, результаты в беге на 100, 800 и 5000 м зависят от скоростных качеств спортсмена, его силы, выносливости и пр. Значение этих факторов для каждой из дистанций различно. Если мы выберем два теста, на которые влияют примерно в равной степени одни и те же факторы, то результаты в этих тестах будут сильно коррелировать друг с другом (скажем, в беге на близких дистанциях, 800 и 1000 м). Если же у тестов нет общих факторов или они мало влияют на результаты, корреляция между этими тестами будет низкой (например, корреляция между результатами в беге на 100 и 5000 м). Когда берется большое число разных тестов и рассчитываются коэффициенты корреляции между ними, то с помощью факторного анализа можно определить, сколько факторов совместно действуют на данные тесты и какова степень их вклада в каждый тест. А затем уже легко выбрать тесты (или их комбинации), которые наиболее точно оценивают уровень отдельных факторов. В этом состоит идея факторной информативности тестов. Рассмотрим на примере конкретного эксперимента, как это делается.

Задача состояла в том, чтобы найти наиболее информативные тесты для оценки общей силовой подготовленности студентов-спортсменов III—I разрядов, занимающихся разными видами спорта. С этой целью было обследовано по 15 тестам 108 человек. (Н. В. Аверкович и В. М. Зациорский, 1966). В результате факторного анализа выделились три фактора: 1) сила верхних конечностей, 2) сила нижних конечностей, 3) сила мышц брюшного пресса и сгибателей тазобедренных суставов. Наиболее информативными тестами среди опробованных оказались: по первому фактору — отжимание в упоре, по второму — прыжок в длину с места, по третьему — поднимание прямых ног в висе и максимальное число подъемов в сед в течение 1 мин. Если ограничиваться лишь одним тестом, то наиболее информативным был переворот силой в упор на перекладине (оценивалось число повторений).

Эмпирическая информативность в практической рабо-

ie. При практическом использовании показателей эмпирической информативности следует иметь в виду, что они справедливы лишь по отношению к тем контингентам испытуемых и условиям, для которых они рассчитаны. Тест, информативный в группе начинающих, может оказаться совершенно неинформативным, если попытаться его применять в группе мастеров спорта.

Одной из причин, которая сильно снижает информативность теста, является отбор испытуемых. Если определена информативность теста на какой-либо группе, а затем сильнейшие из них включены в сборную команду, то информативность того же теста в сборной команде будет значительно ниже. Причины этого понятны из рис. 3:

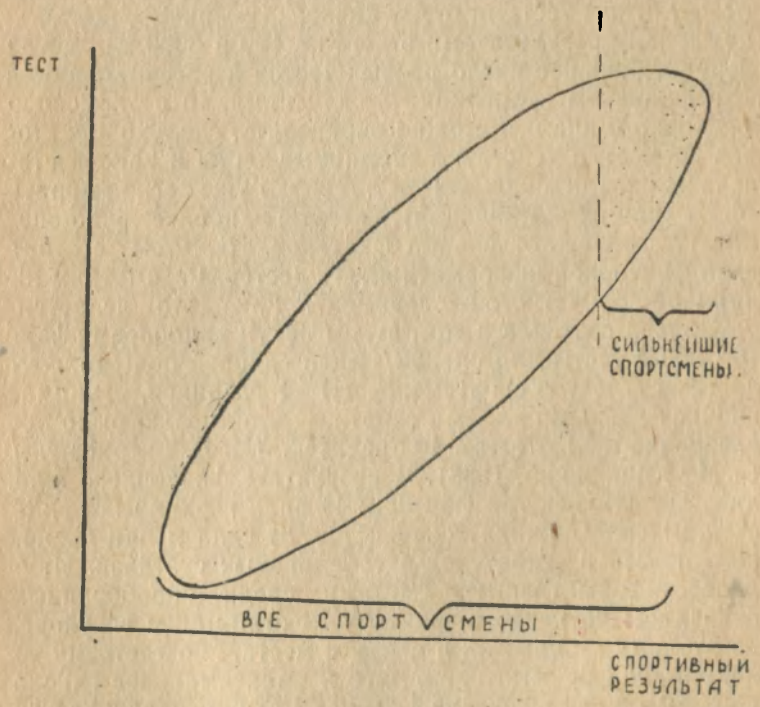


Рис. 3. Влияние отбора испытуемых на информативность теста.

отбор уменьшает общую дисперсию результатов в группе и снижает величины коэффициента корреляции. Например, если определить информативность такого теста, как

МПК, у пловцов на 400 м, имеющих резко различные результаты (скажем, от 3.55,0 до 6.30,0), то коэффициент информативности будет очень высок ($r_{tk}>0{,}90$); если провести те же измерения в группе, где результаты находятся в пределах 3.55,0—4.30,0, то r_{tk} по абсолютной величине не будет превосходить 0,4—0,6; если определить тот же показатель у сильнейших пловцов мира ($3.53{,}0<t<4.00{,}0$), то коэффициент информативности вообще может быть равен нулю; с помощью одного этого теста нельзя будет различить пловцов, проплывающих, скажем, за 3.59,0 и 3.55,0: и у тех, и у других величины МПК будут высоки и примерно одинаковы.

Коэффициенты информативности очень сильно зависят от надежности теста и критерия. Тест с низкой надежностью всегда мало информативен, поэтому малонадежные тесты нет даже смысла проверять на информативность. Если недостаточна надежность критерия, это также приводит к снижению коэффициентов информативности. Однако в данном случае было бы неправильно отбрасывать тест как малоинформативный: вспомним, что верхней границей возможной корреляции является не ± 1, а его индекс надежности. Поэтому надо сравнивать коэффициент информативности с этим индексом. Действительную информативность (с поправкой на ненадежность критерия) рассчитывают по формуле:

$$\hat{r}_{tk} = \frac{r_{tk}}{\sqrt{r_{kk}}} \qquad (7).$$

Так, в одной из работ ранг спортсмена (его рассматривали как критерий мастерства) в водном поло был установлен на основе оценок 4-х экспертов. Надежность (согласованность) критерия, определенная с помощью внутриклассового коэффициента корреляции, равнялась 0,64. Коэффициент информативности был равен 0,56. Действительный коэффициент информативности (с поправкой на ненадежность критерия) равен:

$$\hat{r}_{tk} = \frac{0{,}56}{0{,}64} = 0{,}70$$

С информативностью и надежностью тесно связано понятие о различительной возможности теста, под которой понимается то минимальное различие между испытуемыми, которое диагностируется с помощью теста

(это понятие по смыслу аналогично понятию о чувствительности прибора). Различительная возможность теста зависит от:

1. Межиндивидуальной вариации результатов. Например, такой тест, как «максимальное число повторных бросков баскетбольного мяча в стену с расстояния 4 м в течение 10 сек.», хорош для начинающих, но непригоден для квалифицированных баскетболистов, так как они все начинают показывать примерно один и тот же результат и становятся неразличимыми друг от друга. Во многих случаях вариация результатов между испытуемыми (межклассовая вариация) может быть повышена за счет увеличения трудности теста. Например, если дать спортсменам разной квалификации легкую для них функциональную пробу (скажем, 20 приседаний или работу на велоэргометре мощностью 200 кгм/мин), то величина физиологических сдвигов у всех будет примерно одинакова и оценить степень подготовленности будет невозможно. Если же предложить им трудное задание, то различия между спортсменами станут большими, и по результатам теста окажется возможным судить о подготовленности спортсмена.

2. Надежности (т. е. соотношения меж- и внутриндивидуальной вариации) теста и критерия. Если результаты одного и того же испытуемого в прыжках в длину с места варьируют, скажем, в пределах ± 10 см, то, хотя длину прыжка и можно определить с точностью ± 1 см, различить с убежденностью испытуемых, «истинные» результаты которых равны 315 и 316 см, нельзя.

Нет фиксированной величины информативности теста, после которой можно считать тест пригодным. Здесь многое зависит от конкретной ситуации: желаемой точности прогноза, необходимости получить хоть какие-то дополнительные сведения о спортсмене и т. п. Практически для целей диагностики используются тесты, информативность которых не меньше 0,3. Для прогноза, как правило, нужна более высокая информативность — не менее 0,6.

Информативность батареи тестов, естественно, выше, чем информативность одного теста. Нередко бывает так, что информативность одного отдельно взятого теста слишком низка, чтобы этим тестом пользоваться. Инфор-

мативность же батареи тестов может быть вполне достаточна.

Содержательная (логическая) информативность. Информативность теста не всегда можно установить с помощью эксперимента и математической обработки его результатов. Например, если стоит задача разработать билеты для экзаменов или темы дипломных работ (это ведь тоже разновидность тестирования), надо отобрать такие вопросы, которые наиболее информативны, по которым можно точнее всего оценить знания выпускников и их подготовленность к практической работе. Пока в подобных случаях опираются лишь на логический, содержательный анализ ситуации.

Что касается чисто спортивных тестов, то и здесь бывает, когда их информативность ясна без всяких экспериментов. Чаще всего это случается, когда тест является просто частью тех действий, которые выполняет спортсмен на соревнованиях. Едва ли нужны эксперименты, чтобы доказать информативность таких показателей, как время выполнения поворотов в плавании, скорость на последних шагах разбега в прыжках в длину, процент попаданий со штрафных бросков в баскетболе, качество выполнения подачи в теннисе и волейболе.

Однако не все подобные тесты в равной мере информативны. Например, вбрасывание из-за боковой линии в футболе, хотя и является элементом игры, едва ли может рассматриваться как один из самых важных показателей мастерства футболистов. Если таких тестов становится чересчур много и надо отобрать самые информативные из них, без математических методов теории тестов не обойтись.

Содержательный анализ информативности теста и экспериментально-математическое ее обоснование должны дополнять друг друга. Ни один из этих подходов, взятый сам по себе, не является достаточным. В частности, если в результате эксперимента определен высокий коэффициент информативности, нужно обязательно проверять, не является ли это следствием так называемой ложной корреляции. Напомним, что ложные корреляции появляются, когда на результаты обоих коррелируемых признаков влияет некоторый третий показатель, который сам по себе нас не интересует. Например, у учеников средней школы можно найти существенную корреляцию

между результатом в беге на 100 м и знанием геометрии. Это произойдет потому, что ученики старших классов по сравнению с младшеклассниками в среднем покажут более высокие показатели как в беге, так и в знании геометрии. Посторонним третьим признаком, вызвавшим появление корреляции, явился возраст испытуемых. Конечно, совершил бы ошибку тот исследователь, который этого бы не заметил и рекомендовал экзамен по геометрии как тест для бегунов на 100 м. Чтобы не совершать подобных ошибок, надо обязательно проанализировать причинно-следственные связи, вызвавшие появление корреляции между критерием и тестом. Полезно, в частности, представить себе, что произойдет, если результаты в тесте улучшатся. Приведет ли это к росту результатов критерия? В данном примере это означает: если ученик будет лучше знать геометрию, станет ли он быстрее бежать 100 м? Очевидный отрицательный ответ приводит к естественному заключению — знания по геометрии не могут служить тестом для спринтеров. Найденная корреляция является ложной. Разумеется, ситуации реальной жизни значительно сложнее этого нарочно огрубленного примера.

Частным случаем содержательной информативности тестов является информативность по определению. В данном случае просто договариваются о том, какой смысл надо вкладывать в то или иное слово (термин). Например, если говорят «прыжок в высоту с места характеризует прыгучесть», то точнее было бы сказать так: «условимся называть прыгучестью то, что измеряется результатом прыжка вверх с места». Такой взаимный уговор необходим, так как он предупреждает ненужные недоразумения (ведь кто-то может понимать под прыгучестью результаты в десятерном прыжке на одной ноге, а прыжок в высоту с места считать, скажем, тестом «взрывной» силы ног).

ОСНОВЫ ТЕОРИИ ПЕДАГОГИЧЕСКИХ ОЦЕНОК

Проблема оценок

Основные понятия. Показанные спортсменами результаты (в частности, результаты тестов):

1) выражаются в разных единицах измерения (время, расстояние и т. п.) и поэтому непосредственно не сопоставимы друг с другом;

2) сами по себе не указывают, насколько удовлетворительно состояние спортсмена (скажем, время бега 100 м, равное 12,0, может рассматриваться как очень хорошее или, наоборот, очень плохое, в зависимости от того, о ком идет речь).

Поэтому результаты превращают в оценки (очки, баллы, отметки, разряды и т. п.).

Оценкой (или педагогической оценкой) называется унифицированная мера успеха спортсмена или спортивного коллектива в каком-либо задании, в частном случае о тесте*. Процесс выведения (расчета, определения) оценок называют оцениванием.

Всесоюзная спортивная классификация, комплекс ГТО, таблицы очков по видам спорта, оценки результатов, школьные и вузовские отметки по физическому воспитанию, положения о соревнованиях и утвердившаяся практика неофициального подсчета очков на Олимпийских играх — все это примеры оценивания. Оценка может быть выражена различным способом, например, в виде качественной характеристики («хорошо — удовлетворительно — плохо» или «зачет—незачет»), выставляемой отметки (5 градаций, как в нашей школе—от «единицы» к «пятерке»), набранных очков (например, в многоборье), факте выполнения разрядных норм или норм комплекса ГТО—во всех случаях она имеет общие черты.

Различают учебные оценки, которые выставляет преподаватель своим ученикам по ходу учебного процесса, и квалификационные, под которыми понимают все прочие виды оценок (в частности, результатов официальных соревнований, тестирования и др). Резкой грани между учебными и квалификационными оценками нет, однако процедура квалификационного оценивания, как правило, более сложна. Это связано с большей ответственностью таких оценок и необходимостью обеспечить их максимальную справедливость и полезный эффект.

В полном развернутом виде оценивание происходит в два этапа. На первом показанные спортивные результаты

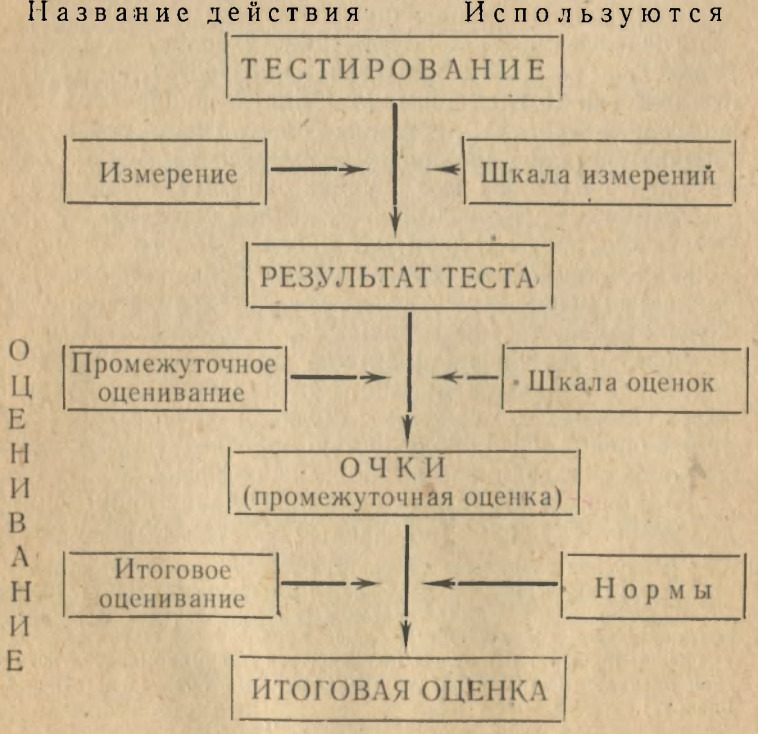
* В научной спортивной литературе термин «оценка» используется в том смысле, как он применяется в математической статистике, — зафиксированное в опыте значение некоторой величины (параметра генеральной совокупности). Когда будет возникать опасность смешения понятий, мы будем пользоваться термином «статистическая оценка» (в противовес определенной выше «педагогической оценке»).

превращают на основе так называемых шкал оценок в очки (промежуточная оценка), а на втором, после сравнения набранных очков с заранее установленными нормами, определяется итоговая оценка. Последовательность действий видна из приводимой ниже схемы, где дополнительно показаны также этапы тестирования и измерения результатов теста.

Например, при присвоении спортивного разряда в многоборьях результаты в отдельных видах переводят в очки (промежуточное оценивание с помощью шкал оценок), а затем после сравнения с установленными нормами спортивной классификации выносится итоговая оценка — присваивается спортивный разряд.

Не во всех случаях оценивание происходит по такой развернутой схеме. Порой промежуточное и итоговое оценивание сливаются вместе.

Схема оценивания спортивных результатов и результатов тестов



Таблицы очков по видам спорта и шкалы оценок. Проанализируем таблицы очков по некоторым видам спорта. Это позволит ввести ряд понятий, необходимых в дальнейшем.

Любая таблица имеет целью преобразование показанного спортивного результата (выраженного в объективных мерах — килограммах, секундах и т. п., в занятом месте или числе и значимости побед) в условные очки. Закон преобразования спортивных результатов в очки

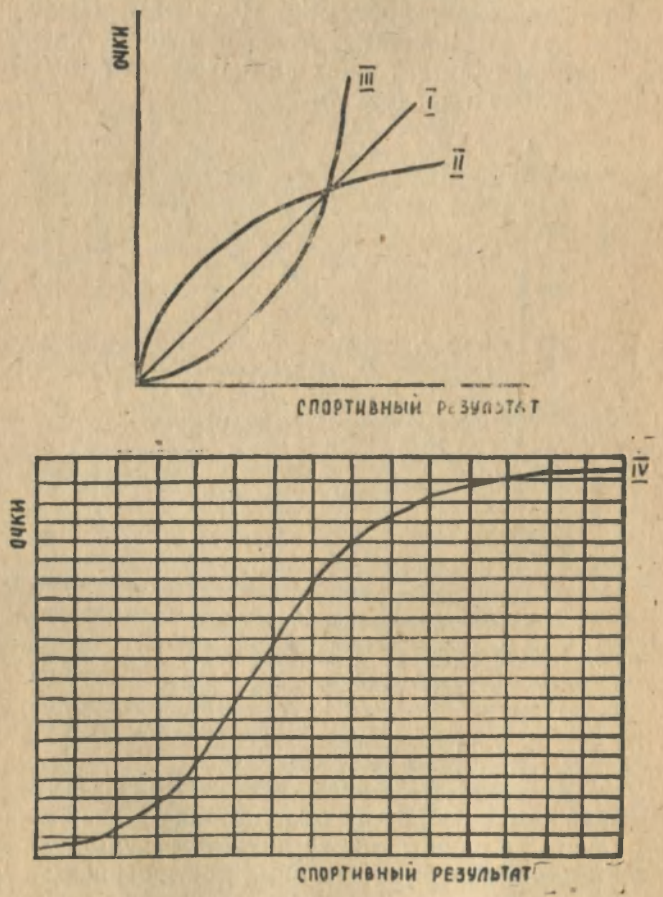


Рис. 4. Основные типы шкал оценок:
1 — пропорциональная шкала, 2 — регрессирующая, 3 — прогрессирующая, 4 — сигмовидная.

называется шкалой оценок. Шкала может быть задана в виде математического выражения (формулы), таблицы или графика. На рис. 4 схематически показаны четыре основных типа шкал, встречающихся в спорте и физическом воспитании. Рассмотрим их.

Первый тип — будем называть его пропорциональной шкалой — предполагает начисление одинакового количества очков за равный прирост результатов (например, за каждые 0,1 с улучшения результата в беге на 100 м спортсмену добавляют 20 очков). Пропорциональные шкалы приняты в современном пятиборье, конькобежном спорте, гонках на лыжах, лыжном двоеборье, биатлоне и др. (рис. 5).

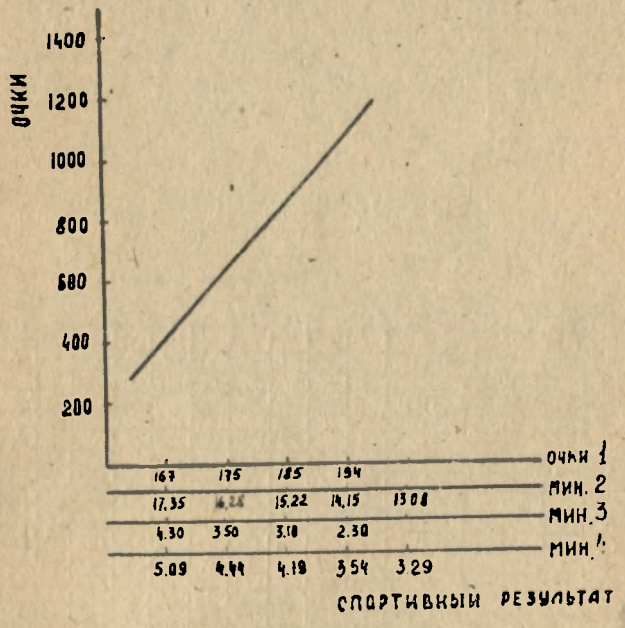


Рис. 5. Шкала оценок результатов в современном пятиборье.

Второй тип — регрессирующие шкалы. В этом случае за один и тот же прирост результатов начисляют с возрастанием спортивных достижений все меньшее количество очков (например, за улучшение результата в беге на 100 м с 15,0 до 14,9 добавляют 20 очков, а за 0,1 в диапазоне от 10,0 до 9,9 — только 15 очков).

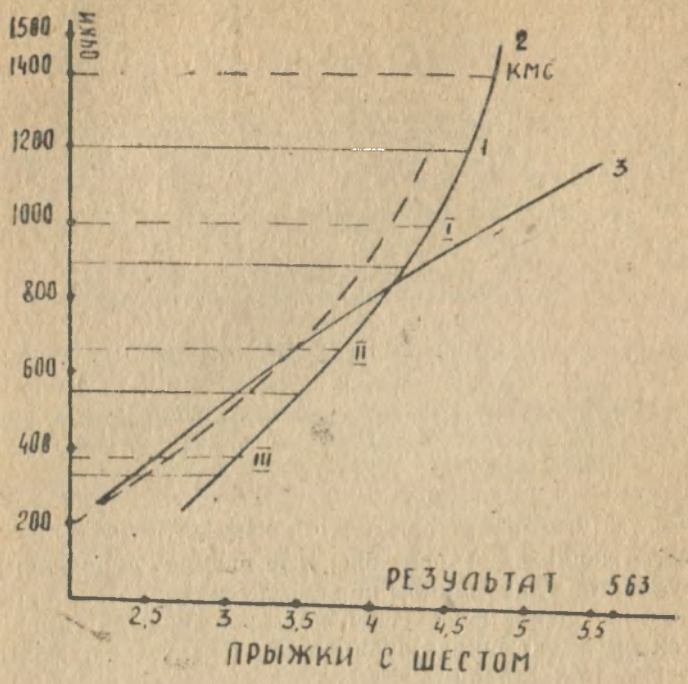


Рис. 6. Шкала оценок результатов в прыжках с шестом (по таблицам очков разных лет).

Такие шкалы обычно кажутся несправедливыми, но они могут быть полезны (раздел «Проблема критерия»). Шкалы такого типа приняты сейчас в легкой атлетике в некоторых видах прыжков и метаний (рис. 6).

Третий тип — прогрессирующие шкалы. Здесь чем выше спортивный результат, тем большей прибавкой очков оценивается его улучшение (например, за улучшение времени бега с 15,0 до 14,9 добавляют 10 очков, а разница между оценками 10,0 и 9,9 составляет, скажем, 100 очков). Прогрессирующие шкалы существуют в плавании, отдельных видах легкой атлетики, тяжелой атлетике и др. (рис. 7).

Четвертый вид — сигмовидные (или S-образные) шкалы. В этих шкалах улучшение результатов в зонах очень низких и очень высоких достижений поощряется скупо; больше всего очков приносит прирост

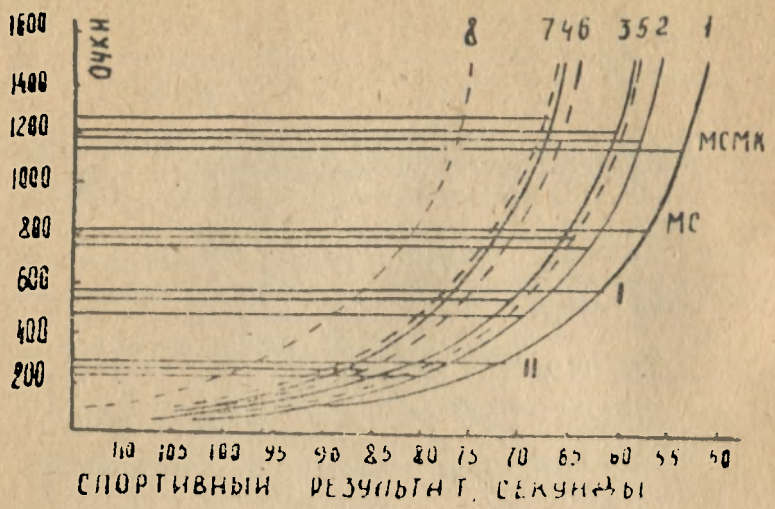


Рис. 7. Шкалы оценок результатов в плавании на 100 м.

достижений в средней зоне. Для оценки спортивных результатов такие шкалы не используются, но они популярны при оценке физической подготовленности (например, так выглядит шкала стандартов физической подготовленности населения США).

Основные задачи оценивания сводятся к необходимости определить:

1. Соответствие между разными достижениями в одном и том же задании (тесте, спортивной дисциплине, упражнении, виде многоборья). Например, сопоставить спортивные результаты, равные норме мастера спорта и I разряда. Сколько перворазрядных результатов соответствует одному мастерскому?

2. Соответствие между достижениями в разных заданиях. Центральным вопросом здесь является уравнивание оценок за достижения равной трудности в разных видах спорта или программы соревнований. Такие равно трудные достижения называются эквивалентными.

3. Нормы. В отдельных случаях (школьные оценки, комплекс ГТО и т. п.) нормы совпадают с градациями шкалы.

Решение этих задач полностью определяет систему оценки.

Проблема критерия. Две группы критериев могут лежать в основе оценки. Оценка должна:

1. Быть справедливой, т. е. оценивать достижения:

а) равной трудности (эквивалентные) — равным количеством очков;

б) неравной трудности — тем большим количеством очков, чем выше их трудность.

2. Приводить к практически полезным результатам.

Эти критерии не всегда совместимы друг с другом. Например, прогрессирующая шкала в принципе представляется справедливой: даже немного повысить мировой рекорд несравненно труднее, чем добиться такого же прироста результатов на уровне III разряда. Эту неравную трудность шкала учитывает: чем выше спортивный результат, тем больше очков начисляется за равный прирост достижений. Практически это приводит к тому, что спортсменам-многоборцам становится выгодно тренировать прежде всего свои любимые виды — те, где они могут получить наибольшее количество очков. В условиях командной борьбы прогрессирующая шкала поднимает ценность высоких спортивных результатов, но чрезмерно высокая прогрессия тормозит массовость: один спортсмен высокой квалификации в этом случае дает намного больше очков команде, чем несколько разрядников.

Регрессирующие шкалы едва ли можно считать справедливыми, но они полезны. В многоборьях они стимулируют внимание к отстающим видам, в командных соревнованиях — массовость (в ущерб мастерству).

Вопрос о том, какая система оценки лучше, бессмыслен, если не поставлена цель, ради которой она вводится. Например, если ставится цель (скажем, на соревнованиях по ОФП) устранить слабые звенья в подготовке, то регрессирующая шкала более приемлема, несмотря на ее несправедливость.

Разумеется, во многих случаях, где это осуществимо, целесообразно сочетать критерии обеих групп («справедливость» и «полезный эффект»).

Уже отмечалось, что непосредственно сопоставлять достижения в разных заданиях нельзя (скажем, не ясно, что труднее — бег 100 м за 11,0 или прыжок в высоту на 2,00 м). Поэтому используют косвенные подходы. Наиболее распространенны шкалы, где эквивалентными счита-

ют достижения, доступные одинаковому числу людей одного пола и возраста. Согласно этому критерию все существующие мировые рекорды эквивалентны и должны оцениваться одинаковым количеством очков, эквивалентны также сотые результаты в списках сильнейших спортсменов, результаты, которые доступны 50% девочек двенадцатилетнего возраста и т. п. Шкалы, основанные на этом критерии, описаны в следующем разделе.

Шкалы оценок

Стандартные шкалы названы так потому, что в них масштабом служат стандартные (средние квадратические) отклонения.

Напомним, что отклонение от средней, выраженное в единицах стандартного отклонения, называется нормированным отклонением. Простейшей стандартной шкалой является Z-шкала. При этом начисляемые очки равны нормированному отклонению. Легко видеть, что в Z-шкале средний результат оценивается в ноль очков, результаты ниже средней получают отрицательные очки, а подавляющее их большинство укладывается в диапазоне от —3,0 до +3,0. Из-за появления отрицательных значений эта шкала не удобна и используется редко.

Наиболее популярна среди стандартных шкал Т-шкала. Здесь средняя приравнивается 50, а стандарт — 10 очкам.

$$T=50+10\times\frac{X-\bar{x}}{\sigma}=50+10Z,$$

где X — показанный результат,

$\bar{x}$ и σ — как обычно, средняя и стандартное отклонение.

Например, если средняя величина прыжков в длину с места равнялась 224 см, а стандарт—200 см, то за результат 222 см начисляется 49 очков, а за 226 см — 72 очка (проверьте, правильно ли это).

Разумеется, приравнивание средней 50, а стандарта 10 очкам произвольно. В мировой практике используются и другие стандартные шкалы.

Некоторые стандартные шкалы

Название	Основная формула	Где и для чего используется
С-шкала	$C=5+2Z$	При массовых обследованиях, когда не требуется большая точность

Шкала школьных отметок	H=3—Z	В ряде стран Европы
Шкала Вине	B=100+16Z	При психологических исследованиях интеллекта
Экзаменационная шкала	E=500+100Z	В США при поступлении в высшие учебные заведения

Стандартные шкалы являются пропорциональными (см. ниже). Они пригодны, если распределение результатов теста близко к нормальному. Используя таблицы нормального распределения, легко узнать, какой про-

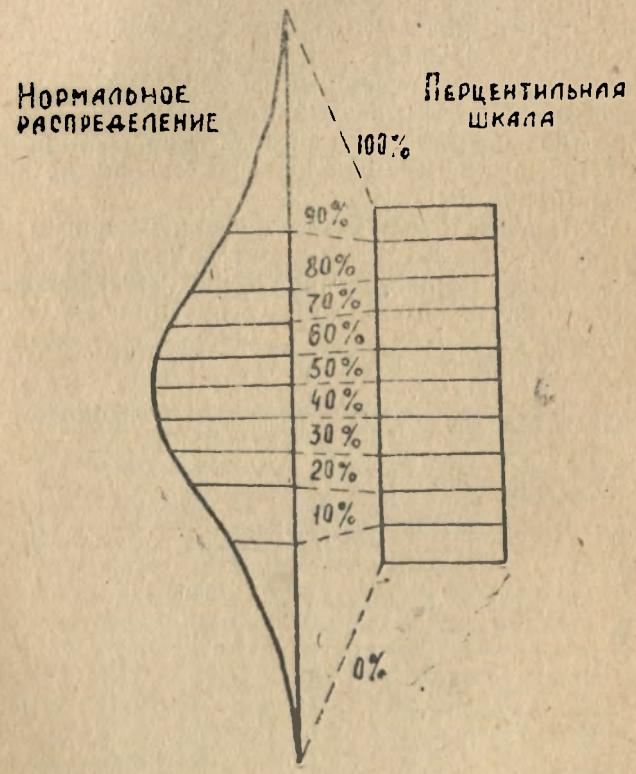


Рис. 8. Соотношение между нормальным распределением и перцентильной шкалой.

цент лиц находится в том или ином диапазоне стандартной шкалы. Например, больше 50 и меньше 60 очков по

Т-шкале будут в среднем набирать 34% всех спортсменов.

Перцентильная шкала. Представим, что проводится кросс с общим стартом. В этом случае можно начислять спортсмену столько очков, какой процент участников он обогнал. Опередил всех (100%) — получает 100 очков, выиграл у 72% — 72 очка и т. д. Тот же принцип можно использовать и в других тестах: число начисляемых очков приравнивать проценту лиц, которых опередил данный участник. Шкала, построенная таким образом, называется перцентильной, а интервал этой шкалы — перцентилем. Один перцентиль включает 1% всех испытуемых, 50%-ный перцентиль, как известно, называется медианой. Поскольку большая часть людей показывает результаты, близкие к средним, и сравнительно мало людей имеет результаты очень высокие или очень низкие, то перцентили соответствуют разным приростам результатов тестов: в середине шкалы — маленьким, на краях — большим (рис. 8).

Перцентильные шкалы относятся к сигмовидным шкалам. Ведь сигмовидные шкалы—это по существу функции нормального распределения (рис. 9). Перцентильные шкалы очень наглядны и поэтому широко используются.

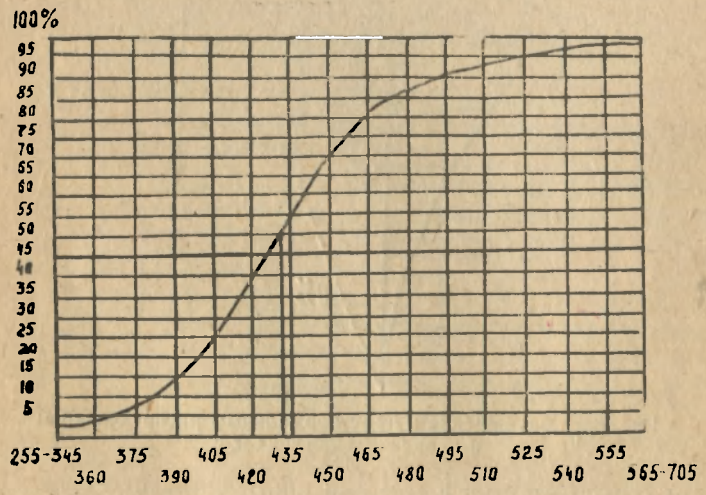


Рис. 9. Пример перцентильной шкалы.

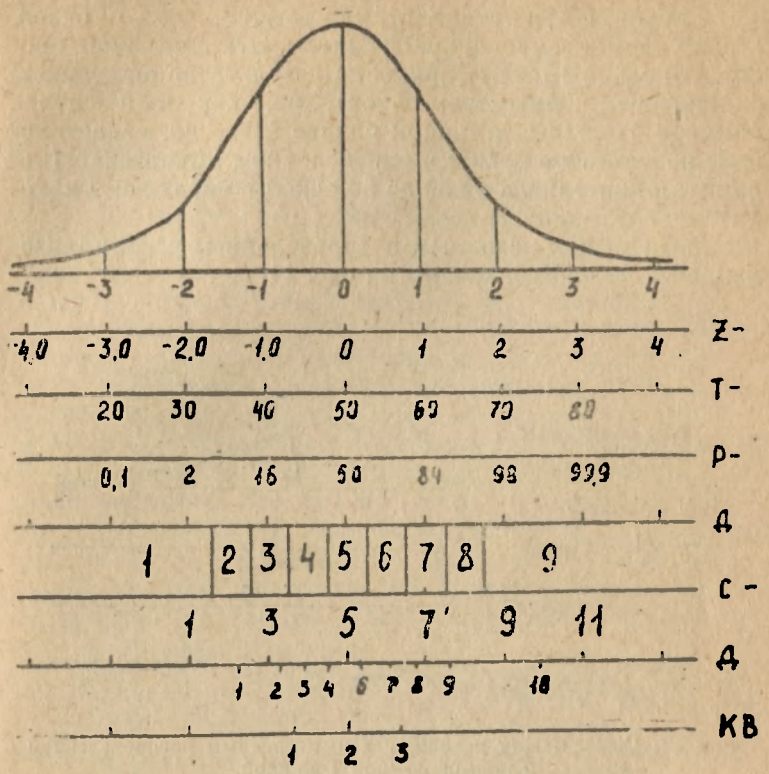


Рис. 10. Наиболее распространенные шкалы и их связь с нормальным распределением.

Рис. 10 иллюстрирует некоторые типы шкал. Мы советуем внимательно изучить этот рисунок.

Нормализованная Т-шкала. В тех случаях, когда распределение результатов теста резко асимметрично, Т-шкалу непосредственно использовать нельзя. Однако ее можно получить путем так называемой нормализации. Это делается в три этапа:

1) находят перцентили исходного асимметричного распределения;

2) используя таблицы нормального распределения, определяют нормированное отклонение, соответствующее данному перцентилю (т. е. поступают так, как будто распределение нормально);

3) принимая медиану распределения (не среднюю) за 50 очков, а один стандарт, как обычно, — за 10 очков, строят нормализированную Т-шкалу. Эта шкала не пропорциональна (равные приросты результатов оцениваются неравным количеством очков), но, как и в случае обычной Т-шкалы, в данном случае известно, какой процент результатов лежит в определенном диапазоне (например; в интервале от 50 до 60 очков по-прежнему находится в среднем 34% всех испытуемых.

Связи между исходным распределением и нормализованной Т-шкалой показаны на рис. 11.

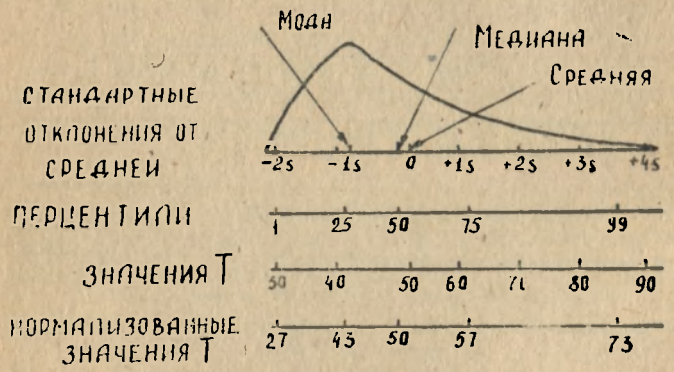


Рис. 11. Связь между исходным асимметричным распределением и нормированной Т-шкалой.

Когда распределение результатов теста нормально, обычная и нормализованная Т-шкала совпадают друг с другом.

Шкалы выбранных точек. Описанные шкалы можно построить, если известно статистическое распределение результатов теста (т. е. средняя, стандарты и другие параметры распределения). Такие данные не всегда можно получить. Это достижимо, например, при разработке шкал для массовых контингентов (комплекс ГТО, нормы по физическому воспитанию в школе и т. п.) и недостижимо при разработке таблиц по видам спорта.

В последнем случае обычно поступают так: берут какой-нибудь высокий спортивный результат (например, мировой рекорд или 10-й результат в истории данного

вида спорта и т. п.) и приравнивают его, скажем, 1000 или 1200 очкам. Затем на основе результатов массовых испытаний определяют среднее достижение группы слабо подготовленных лиц и приравнивают его, скажем, 100 очкам. Если принято решение использовать пропорциональную шкалу, то осталось лишь выполнить арифметические вычисления — ведь две точки однозначно определяют прямую линию. Шкала, построенная таким образом, называется шкалой выбранных точек.

В случае прогрессирующих или регрессирующих шкал возникает сложность с выбором степени их отклонения от прямолинейной зависимости. Например, если за улучшение времени бега с 15,0 до 14,9 начисляется 10 очков, то разница между результатами 10,0 и 9,9 может оцениваться, скажем, в 15 или 150 очков. Обычно такой выбор происходит на основе личных мнений специалистов. Научные методы решения этой задачи не разработаны. В этом, видно, и лежит основная причина того, что почти во всех видах спорта, где используются таблицы очков, многие спортсмены и тренеры не считают их вполне справедливыми.

Параметрические шкалы. В циклических видах спор-

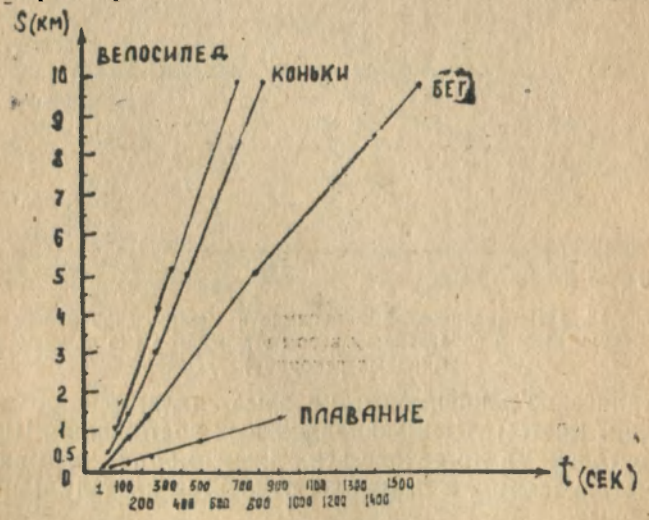


Рис. 12. Параметрическая зависимость между длиной дистанции и временем, по данным мировых рекордов, в циклических видах спорта

та и тяжелой атлетике результаты зависят от таких параметров, как длина дистанций и вес спортсмена. Эти зависимости (спортивного результата от параметра, т. е. длины дистанции или весовой категории) называют параметрическими. Для мировых рекордов они имеют сравнительно простой вид (рис. 12 и 13). Для других эквивалентных достижений (например, равных по трудности II или I разряду) параметрические зависимости должны выглядеть аналогично, т. е. также представлять собой прямые типа (рис. 12 и 13). В принципе можно найта па-

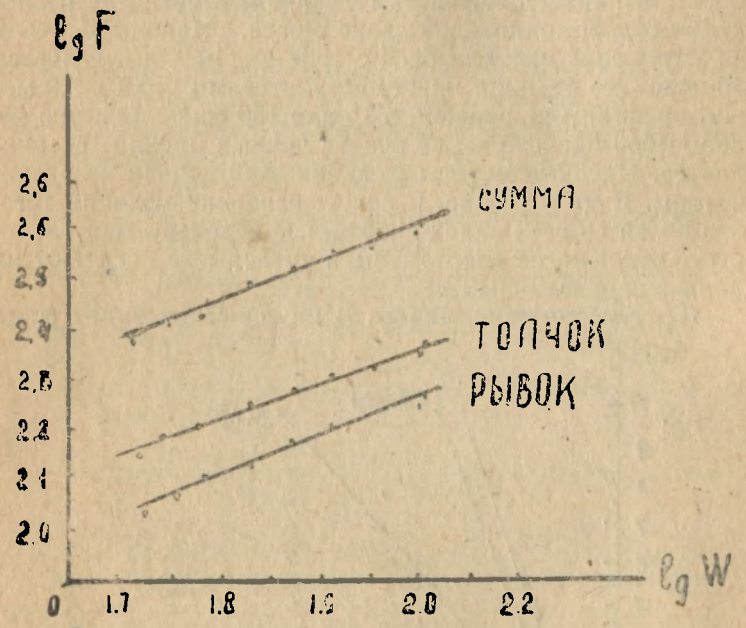


Рис. 13. Параметрическая зависимость между спортивным результатом и собственным весом штангистов (по данным мировых рекордов)

раметрические зависимости, которые являются геометрическим местом точек эквивалентных достижений. Шкалы, построенные на основе этих зависимостей, называют параметрическими. Эти шкалы относятся к числу наиболее точных.

Нормы

Разновидности норм. Нормой в спортивной метроло-

гии называется граничная величина результата, служащая основой для отнесения спортсмена в одну из нескольких классификационных групп. Такими группами могут быть спортивные разряды, градации комплекса ГТО, группы «хорошо» и «недостаточно» тренированных и т. п.

Существует три вида норм:

а) сопоставительные (или популяционные*);

б) индивидуальные;

в) должные.

Сопоставительные нормы имеют в своей основе сравнение людей, принадлежащих к одной и той же популяции. Обычно сопоставительные нормы строятся с помощью шкал, описанных выше в подразделе «Шкалы оценок», но можно вводить нормы и непосредственно на основе средних и стандартов. Например, если вводят 7 классификационных групп, то можно это сделать так, как указано в табл. 5.

Таблица 5

Возможные градации оценок и норм

Оценка		Границы	Процент испытуемых	Нормы в шкалах		
словесная	в баллах			Z	T	перцентильн.
Очень низкая	1	Ниже $x-2\sigma$	2,27	—	—	—
Низкая	2	От $x-2\sigma$ до $x-1\sigma$	13,59	—2,0	30	2,5
Ниже средней	3	От $x-1\sigma$ до $x-0,5\sigma$	14,99	—1,0	40	16,0
Средняя	4	От $x-0,5\sigma$ до $x+0,5\sigma$	38,29	—0,5	45	31,0
Выше средней	5	От $x+0,5\sigma$ до $x+1\sigma$	14,99	+0,5	55	69,0
Высокая	6	От $x-1\sigma$ до $x+2\sigma$	13,59	+1,0	60	84,0
Очень высокая	7	Выше $x+2\sigma$	2,27	+2,0	70	97,5

Примечание. Нормы в перцентильной шкале получаются как округленные суммы процента испытуемых, которым они не доступны.

* От слова «популяция» — группа людей, имеющих общие признаки, например, жители Москвы или двенадцатилетние мальчики, живущие на Кавказе, или спортсмены-баскетболисты II разряда, или легкоатлеты Советского Союза и т. п.

Нормы такого рода удобны тем, что здесь сразу ясно, какому проценту лиц они посильны. Такие нормы целесообразны, когда можно экспериментально зарегистрировать средние значения и стандартные отклонения результатов в той популяции, для которой нормы вводятся.

В сопоставительных нормах используется иногда также другой критерий (кроме процента лиц, которым доступна норма) — время, которое необходимо, чтобы достичь определенного уровня результатов. Например, при определении разрядных норм Всесоюзной спортивной классификации стараются, чтобы сроки подготовки спортсменов одних и тех же разрядов во всех видах спорта были примерно одинаковыми.

Сопоставительные (популяционные) нормы характеризуют лишь сравнительные успехи испытуемых в данной популяции, но они ничего не говорят о популяции в целом. В самом деле, предположим, что в определенном районе в определенных исторических условиях уровень физической подготовленности детей заведомо недостаточен. Если в этом случае построить какую-либо шкалу оценок (например, одну из стандартных шкал) и затем на ее основе ввести нормы, например, так, как это сделано в табл. 5, то заведомо неприемлемый уровень будет признан «средним», и создастся видимость благополучия. Поэтому сопоставительные нормы должны сравниваться с данными, полученными на других популяциях, и использоваться в гибком сочетании с индивидуальными и должными нормами.

Индивидуальные нормы основаны на сравнении показателей одного и того же спортсмена в разных состояниях. Например, во многих видах спорта нет зависимости между собственным весом спортсмена и спортивным результатом: тяжелые и легкие спортсмены могут добиться примерно равных спортивных успехов. Вводить сопоставительную норму здесь не имеет смысла. Однако у каждого спортсмена есть индивидуально-оптимальный вес, соответствующий спортивной форме. Эту индивидуальную норму можно определить, систематически регистрируя вес данного спортсмена в течение достаточно длительного времени. Индивидуальные нормы особенно широко используются в текущем контроле.

Должные нормы основаны на анализе того, что должен уметь делать человек, чтобы успешно справляться с

задачами, которые перед ним ставит жизнь, т. е. труд, оборонная деятельность, быт, спорт и др. Пример: нормы по плаванию в комплексе ГТО было бы неверно вводить на основе среднего уровня умения плавать у людей определенного возраста. Может случиться, что в среднем они плавают недостаточно хорошо. Эти нормы надо вводить с учетом того, как должен уметь плавать человек, чтобы уверенно держаться на воде и преодолевать водные преграды. Здесь, очевидно, целесообразно ввести должную норму.

Таким образом, сопоставительные (популяционные), индивидуальные и должные нормы имеют в своей основе сравнение результатов спортсмена с результатами:

а) других спортсменов;

б) того же спортсмена, но в разные периоды и разных состояниях;

в) должными величинами.

Возрастные нормы. Относятся к сопоставительным. Они основаны на том очевидном факте, что с возрастом функциональные возможности людей изменяются. Есть два варианта возрастных норм. В первом для людей каждого возраста составляется обычным образом одна из шкал оценок (например, перцентильная или Т-шкала) и затем с ее помощью вводятся нормы (скажем, равные 50 или 75 очкам по перцентильной шкале). Во втором определяется так называемый биологический (в частном случае — двигательный) возраст. Он соответствует среднему календарному возрасту людей, которые показывают данный результат. Например, мальчик (неважно какого возраста) прыгнул в длину на 144 см. Средний результат восьмилетних мальчиков равен 140 см

Таблица 6

Двигательный возраст мальчиков по данным прыжков в длину с места

Результат (см)	Двигательный возраст (годы, месяцы)
130	7—1
135	7—6
140	8—0
145	8—5
150	9—1
155	9—9
160	10—8
165	11—8

(табл. 6), а мальчиков возрастом 8 лет 5 месяцев — 145 см. Отсюда легко подсчитать, что 144 см соответствует двигательному возрасту 8 лет 4 месяца (8—4).

Если двигательный возраст опережает календарный, то таких детей называют двигательными акселерантами, если отстает — ретардантами. Например, если три мальчика, одному из которых 7, второму 8, а третьему 9 лет (это их календарный возраст), прыгнули в длину с места на 140 см, то первый из них — акселерант, третий — ретардант, а у второго двигательный возраст по данному тесту соответствует календарному. Из-за неодновременности развития может случиться, что по одним показателям ребенок будет относиться к акселерантам, а по другим — к ретардантам. Полные акселеранты и ретарданты встречаются редко.

При определении возрастных норм людей разделяют на возрастные группы. Например, в комплексе ГТО приняты следующие возрастные группы:

Ступень комплекса	Мужчины	Женщины
I — «Смелые и ловкие»	10—13	10—13
II — «Спортивная смена»	14—16	14—16
III — «Сила и мужество»	16—18	16—18
IV — «Физическое совершенство»	19—39	19—34
V — «Бодрость и здоровье»	40—60	35—55

Видно, что у детей и подростков возрастные градации более частые, чем у взрослых. Это связано с быстрым изменением их двигательных возможностей (табл. 6). В научных исследованиях приняты еще более узкие возрастные группы — не более полугода, а в особо точных случаях — до двух месяцев. Определять возраст в месяцах и днях неудобно. Международные стандарты требуют исчислять возраст по десятичной системе. Для этого следует пользоваться табл. 7.

Возраст при этом определяется так:
возраст (в десятичной системе) = дата тестирования (в десятичной системе) — дата рождения (в десятичной системе).

Например,

дата тестирования: 17 октября 1977 года = 77,792
дата рождения: 20 июля 1961 года = 61,548

Возраст в день тестирования 16,244

Таблица 7

Дни года в десятичной системе

	январь 1	февраль 2	март 3	апрель 4	май 5	июнь 6	июль 7	август 8	сент. 9	октябрь 10	ноябрь 11	декабрь 12
1	000	085	162	247	329	414	496	581	666	748	833	915
2	003	088	164	249	332	416	499	584	668	751	836	918
3	005	090	167	252	334	419	501	586	671	753	838	921
4	008	093	170	255	337	422	504	589	674	756	841	923
5	011	096	173	258	340	425	507	592	677	759	844	926
6	014	099	175	260	342	427	510	595	679	762	847	929
7	016	101	178	263	345	430	512	597	682	764	849	932
8	019	104	181	266	348	433	515	600	685	767	852	934
9	022	107	184	268	351	436	518	603	688	770	855	937
10	025	110	186	271	353	438	521	605	690	773	858	940
11	027	112	189	274	356	441	523	608	693	775	860	942
12	030	115	192	277	359	444	526	611	696	778	863	945
13	033	118	195	279	362	447	529	614	699	781	866	948
14	036	121	197	282	364	449	532	616	701	784	868	951
15	038	123	200	285	367	452	534	619	704	786	871	953
16	041	126	203	288	370	455	537	622	707	789	874	956
17	044	129	205	290	373	458	540	625	710	792	877	959
18	047	132	208	293	375	460	542	627	712	795	879	962
19	049	134	211	296	378	463	545	630	715	797	882	964
20	052	137	214	299	381	466	548	633	718	800	885	967
21	055	140	216	301	384	468	551	636	721	803	888	970
22	058	142	219	304	386	471	553	638	723	805	890	973
23	060	145	222	307	389	474	556	641	726	808	893	975
24	063	148	225	310	392	477	559	644	729	811	896	978
25	066	151	227	312	395	479	562	647	731	814	899	981
26	068	153	230	315	397	482	564	649	734	816	901	984
27	071	156	233	318	400	485	567	652	737	819	904	986
28	074	159	236	321	403	488	570	655	740	822	907	989
29	077		238	323	405	490	573	658	742	825	910	992
30	079		241	326	408	493	575	660	745	827	912	995
31	082		244		411		578	663		830		997

Учет особенностей телосложения. Размеры тела (длина тела, вес и пр.) влияют на двигательные возможности людей. Например, люди высокого роста имеют пре-

имущество в прыжках в высоту. Естественно желание при составлении норм определить их максимально справедливо, чтобы различия в телосложении на них не сказывались.

Наиболее простой путь для этого — выбрать такие тесты, на которые не влияют особенности телосложения. Например, у девочек максимальная скорость бега не зависит от длины тела, а у мальчиков эта зависимость существует только в период полового созревания (рис. 14).

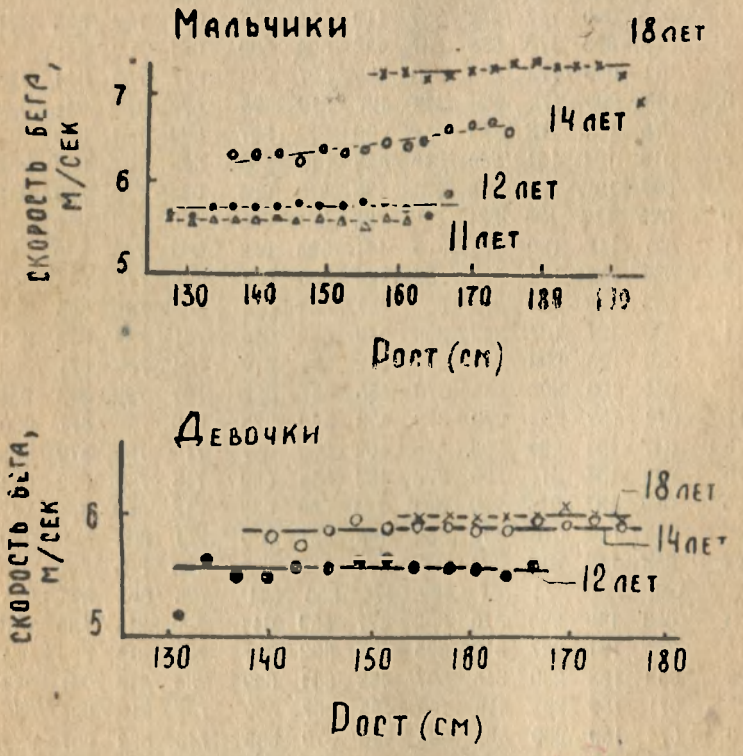


Рис. 14. Максимальная скорость бега у детей разного возраста и длины тела (данные свыше 100 тыс. наблюдений).

Если подобные тесты подобрать не удается, приходится вводить нормы с учетом не только возраста, но также роста и веса. Пример номограмм для определения среднего результата в прыжке в длину с места у 15-летних

мальчиков и девочек приведен на рис. 15. Чтобы определить средний результат, надо соединить на номограмме прямой линией значения роста и веса. Пересечение этой линии со шкалой результатов в прыжке в длину с места укажет среднее значение в этом тесте. Той же цели служат

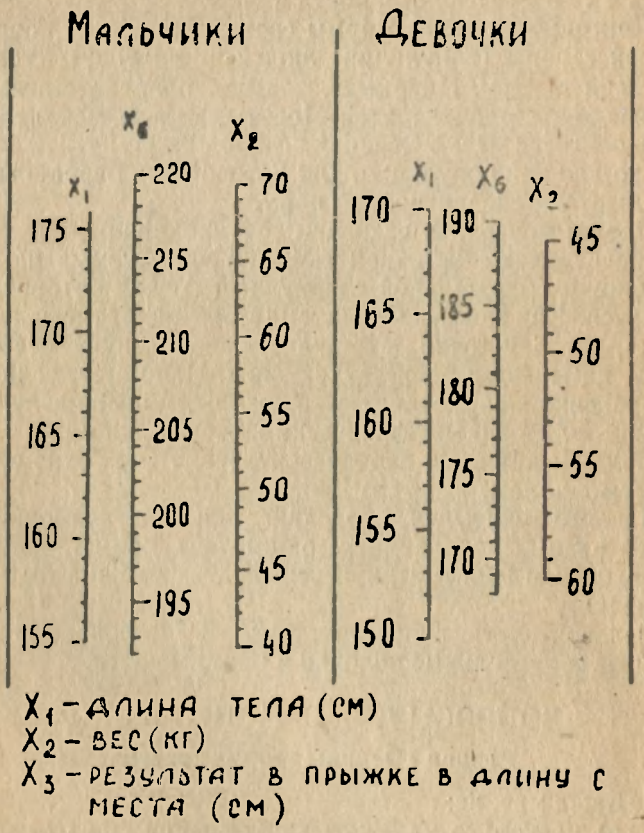


Рис. 15. Номограммы для определения среднего результата в прыжках в длину с места у 15-летних мальчиков и девочек разного роста и веса.
Обратите внимание, что шкалы веса тела у мальчиков и девочек разнонаправленны. Объясните почему?

так называемые классификационные индексы (КИ). Один из них, используемый для оценки физической подготов-

ленности школьников США и Канады, выглядит так:

КИ=20 возраст (в десятичной системе)+2,5 рост(см) +2,0 вес (кг) — 12.

Для каждого значения КИ разработана перцентильная шкала. Определив значение КИ для отдельного испытуемого, можно оценить его физическую подготовленность с учетом возраста, роста и веса.

Пригодность норм. Нормы составляются для определенной группы (популяции) людей и пригодны только для этой группы. Например, нормы, разработанные на основе обследования детей Москвы, нельзя механически переносить на детей Средней Азии. Пригодность норм только для той популяции, для которой они разработаны, называется релевантностью норм.

Нормы пригодны, если они устанавливались на основе обследования типичной выборки испытуемых из всей популяции (генеральной совокупности), для которой они вводятся. Как известно из математической статистики, выборка, которая точно отражает генеральную совокупность, называется репрезентативной. Например, если для определения норм отбираются школы, имеющие лучшие условия для занятий физическим воспитанием, то такая выборка может быть нерепрезентативна по отношению ко всем школам.

Наконец, учитывая, что двигательные возможности людей разных поколений не одинаковы, нормы должны периодически пересматриваться. Норма должна быть современна.

Релевантность, репрезентативность и современность норм — обязательные условия их пригодности.

ВОПРОСЫ ДЛЯ САМОПРОВЕРКИ

Раздел «Основы теории тестов»

1. Ваше мнение:
   а) насколько надежны ваши оценки по математике в аттестате зрелости;
   б) достаточно ли информативны те экзамены, которые вы сдавали при поступлении в институт?
2. Что измеряется при сопоставлении оценок:
   а) по сочинению на выпускных экзаменах в школе с оценками по сочинению при вступительных экзаменах в институт;

б) по одному и тому же сочинению, которые выставили разные преподаватели;

в) оценок при вступительных экзаменах в институт с оценкой за дипломную работу при окончании института?

3. Основные разновидности надежности. Методы ее оценки.

4. Придумайте схему эксперимента, где можно было бы проверить надежность теста и влияние на нее:

а) колебаний в состоянии испытуемых в разные дни недели;

б) тестирования в утренние и вечерние часы;

в) присутствия посторонних наблюдателей.

Сколько раз в таком эксперименте должен тестироваться каждый испытуемый (минимальное значение)?

5. Как определить согласованность оценок судей на соревнованиях по фигурному катанию?

6. Основные разновидности информативности. Методы ее оценки.

7. Ваше мнение — информативны ли:

а) максимальное количество кругов двумя на коне — для оценки специальной выносливости гимнастов;

б) сила сгибателей и разгибателей локтевого сустава— для оценки специальной силовой подготовленности штангистов;

в) подвижность в плечевых, тазобедренных и голеностопных суставах для гимнастов, тяжелоатлетов, копьеметателей;

г) результаты в беге на коньках на 500 м (или 1500, 10.000 м) — для оценки подготовленности в конькобежном многоборье;

д) метание в цель — для отбора в секцию стрельбы из лука;

е) вопросы, на которые вы сейчас ответили, для оценки ваших знаний по теории тестов.

8. Как можно проверить справедливость ответов на вопросы п. 7? Какие эксперименты для этого надо провести?

Раздел «Основы теории педагогических оценок»

1. Справедлива ли система оценок в вашем виде спорта? Обоснуйте свое мнение.

2. Основные типы шкал оценок.

3. Основные разновидности норм.

4. Какие исследования надо провести, чтобы определить шкалы оценок и нормы:

а) комплекса ГТО в беге на 100 м;

б) результатов в плавании и гонках на лыжах для студентов вашего института, не специализирующихся в этих видах;

в) физической подготовленности футболистов команд высшей лиги;

г) всесоюзной спортивной классификации — в плавании, в вашем виде спорта.

5. Какие разновидности шкал и норм вы рекомендуете использовать:

а) при отборе в детские спортивные школы;

б) в общеобразовательной школе;

в) на вступительных экзаменах в институт физкультуры;

г) для оценки подготовленности спортсменов в вашем виде спорта.

ЛИТЕРАТУРА

РАЗДЕЛ «ОСНОВЫ ТЕОРИИ ТЕСТОВ»

Бубэ Х., Фэк Г., Штюблер Х., Трогш Х. Тесты в спортивной практике (перевод с немецкого). ФиС, 1968.

Зациорский В. М., Годик М. А., Ярмульник Д. Н. Теоретические основы и практические пути применения математических методов для оценки специальной физической подготовленности спортсменов. «Теория и практика физической культуры», т. 27, 1964, № 2.

Аверкович Н. В., Зациорский В. М. Факторный анализ тестов силовой подготовленности. «Теория и практика физической культуры», т. 29, 1966, № 8.

Годик М. А., Озолин Э. С., Шустин Б. Н. О корректности измерительных и вычислительных процедур в спортивно-педагогических исследованиях. «Теория и практика физической культуры», 1974, № 3.

РАЗДЕЛ «ОСНОВЫ ТЕОРИИ ПЕДАГОГИЧЕСКИХ ОЦЕНОК»

Бондаревский Я. Е., Парнас В. П., Данилов Ю. Г. К вопросу о статистическом распределении результатов физической подготовленности студентов. «Теория и практика физической культуры», 1975, № 8.

СОДЕРЖАНИЕ

ЛИТЕРАТУРА

Зациорский В. М. Кибернетика, математика, спорт. ФиС, 1969, стр. 12—18, 87—97.

Зациорский В. М., Бондаревский Е. Я., Петросян А. Н. Проблема, оценки спортивных достижений. Методический кабинет ГЦОЛИФКа, М., 1976.

Зациорский В. М., Петров И. Ф. Некоторые практические аспекты анализа зависимости между силой и собственным весом спортсмена. «Теория и практика физической культуры», 1964, № 7.

Обл. тип. им. М. Горького Костромского управления издательств, полиграфии и книжной торговли, зак. 4728, тир. 1000.