К проблеме статистики школьных отметок

Х. Ф. Таммет

Оценка знаний учащихся всегда была в числе сложнейших проблем школьной практики и педагогической науки, вызывающих дискуссии, порождающих различные, нередко диаметрально противоположные предложения о путях их решения. В данной статье мы остановимся лишь на одном вопросе этой многоаспектной проблемы: правомерно ли рассматривать школьные отметки как числа и делать содержательные выводы на основе арифметических средних и коэффициентов корреляции, вычисленных по ним? Вопрос этот спорный, о чем свидетельствуют расхождения между теорией и практикой. В практической работе средний балл широко используется на правах содержательной величины. Теоретики же склонны решительно отрицать правомерность любых арифметических операций с отметками. Их точка зрения освещена в статьях Л. М. Фридмана (Советская педагогика, 1971, № 3) и А. А. Пинского (Советская педагогика, 1980, № 12). В последней из них сказано: «Вместе с тем введение цифровых баллов породило иллюзию, будто бы баллы — это числа, над которыми можно проводить арифметические действия, в частности вычислять средний балл. Очевидно, что это ошибка: цифровые баллы эквивалентны словесным оценкам и образуют ранговую (порядковую) шкалу...».

В сложившейся ситуации возможны два решения: или решительно отказаться от арифметических действий с отметками или найти теоретическое обоснование для арифметических действий с отметками и лишь по мере необходимости уточнить соблюдаемые в практике правила. В настоящей статье приводятся некоторые соображения, свидетельствующие о том, что возможен и второй путь. *Процесс определения отметки* — частный случай процесса измерения, поэтому в наших рассуждениях мы постараемся учитывать основные положения общей теории измерения, изложенной, в частности, в работах: Суппес П., Зинес Дж. Основы теории измерений. — В кн.: Психологические измерения. М., 1967; Пфанцагл И. Теория измерений. М., 1976. Поскольку ниже рассматривается только определенный частный вид измерения, то используется и соответствующая частная терминология. Основные понятия будут отмечены краткими условными терминами: ответ — рассматриваемое действие рассматриваемого индивида или протокол этого действия, отметка — окончательный результат измерения. Школьная отметка имеет пять возможных значений. В большинстве случаев измеряемое свойство ответа имеет больше градаций. Это выражено, например, в указаниях по оценке знаний и умений учащихся по математике в средней школе, где сказано следующее: «Устный ответ и письменная контрольная работа при условии безупречного выполнения всех заданий оценивается отметкой «5». Если выполнена лишь часть работы, но безупречно, то выставляется другая оценка, число баллов которой соответствует выполненной части работы. За допущенные при устном ответе или в контрольной работе ошибки или недочеты оценка снижается на один балл за ошибку и на треть балла за недочет и выставляется ближайшая к полученному результату отметка».

На основе сказанного мы рассмотрим определение отметки как двухэтапный процесс: ответ→предотметка→отметка, где предотметка может выражаться любым дробным числом. Договоримся, что предотметки и отметки согласованы таким образом, что в случае целочисленной предотметки отметка и предотметка равны. Переход от предотметки к отметке рассматривается как округление, которое должно уменьшить нежелательные следствия мелких случайных погрешностей определения предотметки. Цитированные выше указания из програмговой структуры шкалы измерения не исключает возможности существования интервальной структуры той же шкалы. Когда необходимо подчеркнуть, что шкала измерения имеет ранговую и не имеет интервальной структуры, мы будем пользоваться термином «собственно ранговая» шкала.

Деформации шкалы измерения. Одну шкалу измерения можно преобразовать в другую путем замены отметок a на новые отметки a' согласно некоторому конкретному правилу $a'=f(a)$. Мы называем деформациями шкалы измерения такие преобразования, которые описываются монотонно возрастающими функциями. Каждая деформация $a'=f(a)$ имеет обратное преобразование $a=f^{-1}(a')$, которое является также деформацией и называется обратной деформацией относительно первой. Деформацию можно описать при помощи таблицы, математического уравнения или графика. В качестве конкретного примера рассмотрим простую деформацию, которая описывается уравнением $a'=(a^2+5)/6$, таблицей

Старая предотметка	a	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
Новая предотметка	a'	1,0	1,2	1,5	1,9	2,3	2,9	3,5	4,2	5,0

или графиком в виде параболы.

Ранговая структура шкалы измерения сохраняется при любых деформациях, а интервальная только при линейных. Нелинейная деформация преобразует интервальную шкалу в собственно ранговую. Очевидно, соответствующая обратная деформация способна преобразовать некоторую собственно ранговую шкалу в интервальную шкалу измерения. Последнее преобразование называется метризацией ранговой шкалы измерения. Не всякая ранговая шкала допускает метризацию: она возможна только в случае существования во множестве ответов внутренней интервальной структуры, которая в первичной шкале измерения по какой-то причине не используется.

Интервальные шкалы, связанные друг с другом посредством линейной деформации, называются ниже эквивалентными интервальными шкалами. Неэквивалентные интервальные шкалы измерения одной и той же величины могут существовать только в случае, когда во множестве ответов независимо определены разные самостоятельные интервальные структуры. Однако не исключено, что интервальные шкалы, построенные на основе независимо определенных внутренних структур множества ответов, могут оказаться эквивалентными друг другу.

Операция осреднения ответов и интервальная шкала измерения. Выше было сказано, что интервальную структуру множества ответов можно установить косвенным путем при наличии эмпирического правила сравнения интервалов на эквивалентность. Такое правило, в свою очередь, может быть построено косвенным путем, при котором непосредственно определяется операция осреднения ответов. Отметим эквивалентность двух ответов A и B выражением $A=B$ и обозначим средний относительно A и C ответ через $\overline{AC}$. Средний ответ должен быть установлен эмпирическим путем, и операция осреднения должна обладать следующими обязательными свойствами: 1) $\overline{AA}=A$; 2) $\overline{AC}=\overline{BC}$ равносильно $A=B$; 3) $\overline{AB}=\overline{BA}$; 4) $\overline{\overline{AB}\,\overline{CD}}=\overline{\overline{AD}\,\overline{CB}}$. Отношение эквивалентности интервалов может быть определено следующим правилом: $(AB)=(CD)$ равносильно $\overline{AD}=\overline{CB}$. Докажем транзитивность этого отношения. Предположим, что $(AB)=(CD)$ и $(CD)=(EF)$. Согласно определению, это равносильно $\overline{AD}=\overline{CB}$ и $\overline{CF}=\overline{ED}$. Составим теперь выражение $\overline{AF}\,\overline{CD}$ и преобразуем его с помощью четвертого и третьего свойства операции осреднения, а также принятого выше предположения: $\overline{\overline{AF}\,\overline{CD}}=\overline{\overline{AD}\,\overline{CF}}=\overline{\overline{CB}\,\overline{ED}}=\overline{\overline{CD}\,\overline{EB}}=\overline{\overline{EB}\,\overline{CD}}$. Отсюда второе свойство операции осреднения позволяет заключить: $\overline{AF}=\overline{EB}$, откуда, согласно, определению, вытекает $(AB)=(EF)$, что и требовалось доказать. Из изложенного вытекает, что умение эмпирическим путем найти средний ответ для любых двух заданных ответов — вполне достаточная основа для построения интервальной шкалы измерения.

Статистическая метризация шкалы измерения. Если во множестве ответов, которое обладает ранговой структурой, определена вероятность $p(AB)$ попадания случайно выбранного ответа в промежуток AB, то это может быть достаточной основой для построения интервальной шкалы измерения. Интервалы (AB) и (CD) можно признать эквивалентными при условии $p(AB)=p(CD)$. Отметки выставляются так, чтобы $в-a=\gamma p(AB)$, где γ—произвольная постоянная. Построенная описанным способом интервальная шкала измерения называется ниже спирменовской шкалой. Как известно, Спирмен изучал проблему описания статистической связи между двумя ранжированными величинами и предложил при вычислении коэффициента корреляции прямые значения величин заменить на ранги этих значений в вариационном ряду. Спирменовская шкала эквивалентна интервальной, построенной на основе операции осреднения, при условии, что за средний ответ $\overline{AB}$ принимается медиана промежутка AB. Вероятностное распределение предотметок в спирменовской шкале равномерно, а округленные постоянным шагом отметки равновероятны.

Стандартная гауссовская шкала получается из спирменовской шкалы в результате следующей нелинейной деформации: $a'=\psi\left(\frac{a-a_0}{\gamma}\right)$ где a_0 — минимальная спирменовская предотметка и ψ — обратная функция нормального распределения. Нестандартные гауссовские шкалы получаются из стандартной гауссовской шкалы путем линейных деформаций. Гауссовские предотметки известны в литературе как нормальные *скоры* или нормальные метки. Эквивалентность интервалов устанавливается в гауссовской шкале искусственным способом: $(AB)=(CD)$ равносильно $в'-a'=d'-c'$. Вероятностное распределение гауссовских предотметок всегда является гауссовским нормальным распределением. Как спирменовские, так и гауссовские предотметки инвариантны относительно любых предварительных деформаций исходной ранговой шкалы.

Является ли школьная шкала оценивания ранговой шкалой? Ранговая структура множества ответов обычно предполагается как что-то само собой разумеющееся. В действительности проблема не тривиальна. Рассмотрим пример, где сравнение двух ответов имеет исключительно объективный характер. На уроке физкультуры оценивается умение учащихся провести определенную спортивную игру друг против друга. Если ученик A побеждает ученика B, то ответ B считается хуже ответа A. Возможна ситуация, в которой A побеждает B; B побеждает C, но C побеждает A. В описанном примере мы имеет следующие альтернативы: или отказаться от построения ранговой шкалы измерения, или усовершенствовать правила оценивания так, чтобы требования к ранговой структуре множества ответов были выполнены. В практике всегда выбирается вторая альтернатива. Например, в вышеописанной ситуации устраивается турнир и ответы упорядочиваются по накопленным очкам.

Другой вопрос — воспринимается ли шкала оценивания в школьной практике как собственно ранговая шкала? Здесь уместен пример сравнения составных ответов. Пусть ученики A и B выполнили контрольную работу, состоящую из двух равнозначных задач, и получили следующие предотметки:

	A	B
Первая задача	1,0	1,5
Вторая задача	5,0	1,8

Если контрольная работа одного ученика кажется нам лучше работы другого, то это значит, что мы не воспринимаем шкалу оценивания как собственно ранговую. Для большей наглядности совершим следующую деформацию шкалы:

Старая предотметка	1,0	1,2	1,5	1,8	2,0	...	5,0
Новая предотметка	1,0	3,6	4,7	4,8	4,85	...	5,0

Новые предотметки для вышерассмотренной контрольной следующие:

	A	*B*
Новая задача	1,0	4,7
Вторая задача	5,0	4,8

В случае собственно ранговой шкалы сводный вывод из новых предотметок должен совпадать с выводом из старых предотметок.

Возможно ли добиться интервальной структуры школьной шкалы оценивания? Для построения интервальной шкалы измерения необходима интервальная внутренняя структура множества ответов. Выше были рассмотрены следующие способы установления интервальной структуры: определение эмпирической процедуры непосредственного сравнения интервалов между ответами, определение эмпирической операции осреднения ответов, статистическая метризация предварительно построенной ранговой шкалы.

В указаниях об оценке знаний и умений учащихся по математике в средней школе сказано, что оценка снижается на фиксированную величину за определенную ошибку или недочет независимо от начального уровня: интервал между отметками «3» и «4» измеряется тремя недочетами, равно как и интервал между «4» и «5». Это может привести к прямому построению интервальной шкалы измерения с использованием первого из трех вышеперечисленных способов. Однако такой путь, по-видимому, не всегда успешен. Вторая возможность построения интервальной шкалы кажется нам более универсальной. Рассмотрим пример. Допустим, ученики дадут ответы, состоящие из двух равнозначимых разделов, причем предотметки за разделы ответа в произвольной ранговой шкале — (a', a'') для ученика A и ($в'$, $в''$) для ученика B. Пусть $в'=в''=в$, тогда естественная сводная оценка ученика B будет $в$. Если учитель способен ранжировать составные ответы, то он может для заданного A подобрать такого B, чтобы (a', a'') не хуже и не лучше $в$. В случае соблюдения соответствующих алгебраических требований это и есть операция осреднения, которую можно положить в основу интервальной шкалы оценивания. Использование в педагогических исследованиях таких методов, как спирменовский корреляционный анализ, критерий Ван дер Вардена и т. д., равносильно статистической метризации школьной шкалы оценивания.

Вывод о том, что интервальная шкала оценивания может быть построена разными способами, ставит новую проблему. Какой из них предпочесть? Наиболее благоприятное решение получилось бы, если бы способы привели к эквивалентным интервальным шкалам. Ответ на поставленный вопрос могут дать только специальные экспериментальные исследования в школе.

Средний балл. Обычно считается, что в случае интервальной шкалы измерения средний балл является содержательной величиной, а в собственно ранговой школе он лишен смысла. Это утверждение отражает лишь самые общие черты действительного положения, поэтому желательны некоторые уточнения. Во-первых, и в случае интервальной шкалы измерения нельзя вслепую доверять простому арифметическому среднему отметок $\left(\bar{x}=\sum_{i=1}^{n} x_i/n\right)$, которое интерпретируется как сводная отметка только при условии равнозначимости осредняемых отметок. Однако это условие не всегда выполняется. Например, аттестационный балл вступающего в вуз выпускника школы вычисляется как арифметическое среднее явно неравнозначимых отметок по разным предметам. В случае неравнозначимых отметок при интервальной шкале измерения в качестве сводной отметки может быть рассмотрено взвешенное арифметическое среднее $\bar{x} = \sum_{i=1}^{n} p_i x_i \Big/ \sum_{i=1}^{n} p_i$ где вес p_i измеряет значимость соответствующей отметки. Например, вес предмета может быть приравнен к числу уроков, отведенных на данный предмет. Во-вторых, арифметическое среднее может быть иногда использовано и в случае собственно ранговой шкалы измерения. В этом случае все содержательные выводы должны быть инвариантны относительно любой деформации шкалы. Поскольку расположение как простого, так и взвешенного среднего балла среди отдельных отметок изменяется при деформации шкалы, то средний балл непосредственно не может служить основой содержательных выводов. Арифметическое среднее следует заменить некоторой другой характеристикой центра распределения, которая сохранила бы свое расположение среди отметок при любых деформациях шкалы измерения. Такой характеристикой является медиана вероятностного распределения отметок. Оптимальный в таком смысле способ зависит от закона вероятностного распределения отметок. Оказывается, что в случае гауссовского распределения предотметок оптимальной статистической оценкой медианы является арифметическое среднее предотметок. Это придает среднему баллу содержательный смысл независимо от типа шкалы измерения.

В школьной практике типичная задача: задана выборка отметок, а спрашивается медиана предотметок. При осреднении случайные ошибки, как известно, подавляются и в случае сводных оценок пятибалльная шкала явно недостаточна. Если вероятностное распределение предотметок близко к гауссовскому распределению, то арифметическое среднее отметок может быть рассмотрено как простейшая статистическая оценка медианы предотметок.

Из изложенного вытекают три основных вывода: 1. Шкала оценивания знаний и умений учащихся удовлетворяет требованиям, предъявляемым к ранговым шкалам, только благодаря специальному приспосабливанию правил оценивания; 2. Возможно и такое приспосабливание правил оценивания, которое обеспечит интервальную структуру шкалы измерения. Соответствующие требования в некоторой мере учитываются в школьной практике и отражаются в официальных указаниях относительно оценки знаний и умений учащихся. В той же мере оправдана и практика вычисления среднего балла и других арифметических функций отметок; 3. Существует несколько разных способов обеспечения интервальной структуры шкалы оценивания, и поэтому возникает вопрос, приведут ли разные способы построения интервальной шкалы оценивания к эквивалентным шкалам.

Приведенные рассуждения о характере шкалы оценивания знаний и умений учащихся не решают встречающиеся в практике проблемы. Наша задача — уточнить их постановку.