Машин В.А.

Методическое руководство по оценке тестовых процедур, применяемых при работе с персоналом

(2008)

Содержание

ВВЕДЕНИЕ 4
НАДЕЖНОСТЬ ТЕСТА 5
Общие понятия 5
- Основные факторы, влияющие на надежность тестовых измерений 5
- Классификация основных типов надежности теста 7

Влияние инструмента оценки на надежность измерения 10
Временная устойчивость результатов теста 10
- Ретестовая надежность 10
- Надежность параллельных форм теста 12
- Коэффициент ассоциации Пирсона 14
- Коэффициент корреляции Гилфорда и каппа-коэффициент 16
- Тетрахорический коэффициент корреляции Пирсона 17

Согласованность (однородность) содержания инструмента оценки 18
- Метод расщепления 18
- Формула Спирмена-Брауна 19
- Формула Фланагана 20
- Формула Кристофа 20
- Формула Рюлона 20
- Формула Спирмена-Брауна 21
- Метод Кьюдера-Ричардсона 22
- Формула Гуликсена 23
- Формула KR-20 23
- Коэффициент Кронбаха 24
- Формула Спирмена-Брауна для оценки надежность теста с изменением его длины 25
- Метод дисперсионного анализа 26

Согласованность в скорости выполнения тестовых заданий 26
Независимость оценок от различий между оценщиками 27
- Ошибки оценщика (эксперта) 28
- Основные меры по снижению ошибок оценщика (эксперта) 29
- Поведенчески выверенные оценочные шкалы (BARS) 30
- Вынужденное распределение 32
- Оценка компетентности экспертов (оценщиков) 33
- Межэкспертная надежность - коэффициент конкордации W 35
- Согласованность оценок - коэффициенты корреляций результатов оценщиков 37

Требования к выборке испытуемых при изучении надежности 38
- Количественные требования 38
- Содержательные требования 39

Общий обзор типов и коэффициентов надежности 40
Общие принципы для интерпретации коэффициентов надежности 41
ВАЛИДНОСТЬ ТЕСТА 47
Общие понятия 47
- Классификация основных типов валидности теста 47
- Основные факторы, влияющие на валидность теста 48

Валидизация по предмету измерения (Валидность по содержанию) 49
- Очевидная (внешняя) валидность 49
- Содержательная валидность 50

Валидизация по цели измерения (Валидность по критерию) 52
- Основные типы критериев 53
- Основные требования к критериям 57
- Прогностическая валидность 58
- Текущая валидность 61
- Инкрементная валидность 62
- Конвергентная и дискриминантная валидность 64

Комплексная валидизация 67
- Конструктная валидность 67

Требования к выборке испытуемых при изучении валидности 72
Общие принципы для интерпретации коэффициентов валидности 73
- Форма связи теста и критерия 73
- Стандартная ошибка оценки 75
- Коррекция валидности от значений надежности теста и критерия 77
- Простая стратегия отбора 77

- Отношение валидности к продуктивности. 80
- Понятие полезности в теории принятия решений 83
- Последовательные стратегии и адаптивный подход 84
- Объединение данных различных тестов 86

ДИСКРИМИНАТИВНОСТЬ ТЕСТА 90

- Коэффициент Фергюсона 90
- Коэффициент дискриминации – бисериальный коэффициент корреляции 91
- Индекс дискриминации (D) 93
- Четырехпольный коэффициент корреляции 94
- Коэффициент корреляции Гилфорда – дискриминативность теста 95
- Коэффициент корреляции Гилфорда – дискриминативность заданий 96
- Алгоритм раздельного коррелирования ответов 98

ЛИТЕРАТУРА 100

ПРИЛОЖЕНИЯ 102

1. Коэффициент корреляции Пирсона 102
2. Коэффициент ранговой корреляции Спирмена 105
3. Коэффициент ранговой корреляции Кендалла 106
4. Значимость коэффициентов корреляции 108
5. Анализ работы (содержательная валидность) 109
   - Метод критических случаев 116
   - Интервью с целью анализа работы 118
   - Контрольные листы 122
   - Экспертные группы 124
   - Репертуарные решетки 125
   - Иерархический анализ заданий 127
   - Наблюдение 128
   - Самоописание 128
   - Дневники и рабочие журналы 128
   - Общие процедуры анализа содержания работы 129
6. Таблицы критических значений коэффициента конкордации W 137

ВВЕДЕНИЕ

В настоящее время можно выделить следующие основные психометрические требования, благодаря которым возможно обеспечить достаточно объективный характер тестовых процедур как измерительных методов и эффективность достижения с помощью тестовых процедур поставленных целей:

1) надежность;
2) валидность;
3) дискриминативность.

Даже если методика имеет уже оценки надежности, валидности и дискриминативности, очень часто практика применения теста на новой выборке требует от пользователя коррекции основных психометрических критериев объективности и эффективности тестовой процедуры. Диапазон "срабатывания" многих тестов довольно узок и фактически сводится к той популяции, на которой происходила эмпирико-статистическая разработка теста, обеспечивающая его надежность, валидность и дискриминативность. Следовательно, для корректного применения теста на новой популяции или в новых условиях, от разработчика требуется перепроверка его надежности, валидности и дискриминативности. При разработке нового тестового инструментария вопросы надежности, валидности и дискриминативности становятся центральными.

В данном руководстве главный акцент сделан на практических методах и процедурах достижения и оценки надежности, валидности и дискриминативности тестовых материалов, применяемых в отборе и психологическом сопровождении персонала, при определении эффективности его деятельности. При этом теоретические вопросы помогают раскрыть содержание основных психометрических показателей тестовых процедур.

НАДЕЖНОСТЬ ТЕСТА

ОБЩИЕ ПОНЯТИЯ

Результат психологического исследования (оценка свойств обследуемых) обычно подвержен влиянию большого количества неучитываемых факторов (см. рис. 1). Любое изменение ситуации исследования усиливает влияние одних и ослабляет воздействие других факторов на результат теста.

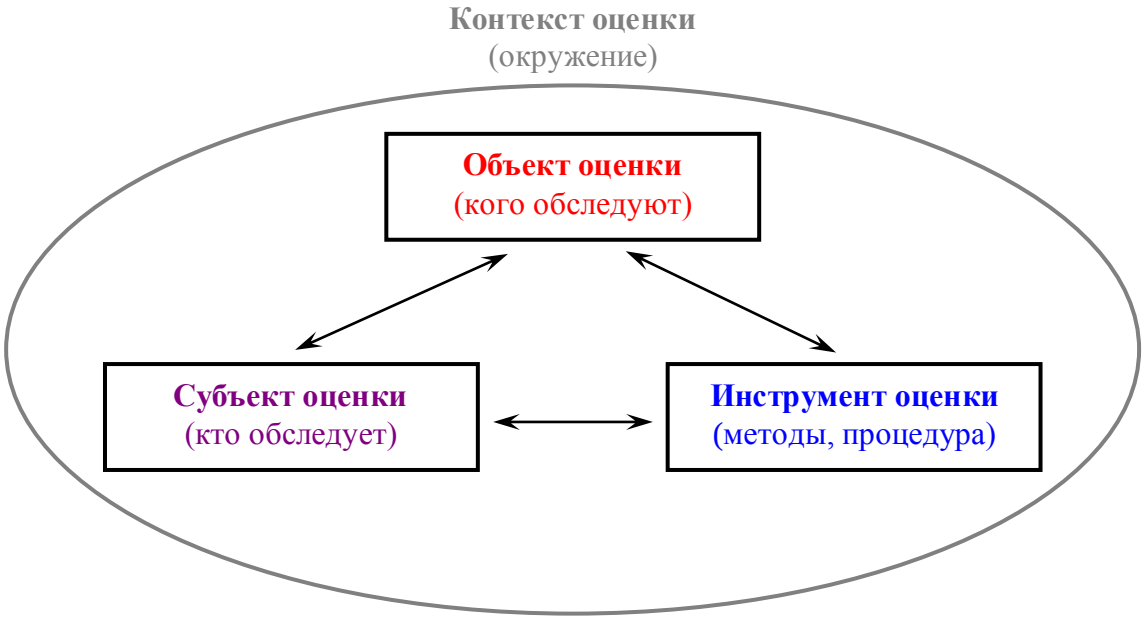


Рисунок 1. Факторы, влияющие на точность (надежность) психологических измерений.

- **Основные факторы, влияющие на надежность тестовых измерений**

Перечислим ряд центральных факторов, влияющих на точность (надежность) измерений психологических свойств и характеристик:

1. Объект оценки

- Нестабильность диагностируемого свойства (например, колебания в уровне компетенции исполнителя).
- Колебания в функциональном состоянии испытуемого (эмоциональное возбуждение, перенапряжение, скука; например, в одном эксперименте отмечается хорошее самочувствие, в другом - утомление) и в его уровне мотивированности на

обследование. Эти колебания могут быть вызваны внешним событием, произошедшим между первым и вторым экспериментом.

- Эффект первого тестирования может влиять на результаты повторного тестирования: например, смена стратегии выполнения.
- Угадывание ответов испытуемыми (главным образом в заданиях с ответами типа "истинно-ложно")[1].

2. Субъект оценки

- Различия в манере поведения экспериментатора (от опыта к опыту по-разному предъявляет инструкции, по-разному стимулирует выполнение заданий и т.д.).

3. Инструмент оценки

- Несовершенство диагностических методик (небрежно составлена инструкция, задания по своему характеру разнородны, нечетко сформулированы указания по предъявлению методики испытуемым и т.д.).
- Элементы субъективности в способах оценки и интерпретации результатов (когда ведется протоколирование ответов испытуемых, оцениваются ответы по степени полноты, оригинальности и т.п.), включая экспертные оценки.

4. Контекст оценки

- Меняющаяся ситуация обследования (разное время дня, когда проводятся эксперименты, разная освещенность, температура и другие особенности помещения, наличие или отсутствие посторонних шумов, разные погодные условия и т.д.).

Общий разброс (дисперсию) результатов тестового обследования можно, таким образом, представить как результат влияния двух групп причин (см. рис. 2): изменчивости, присущей самому измеряемому свойству ("истинная дисперсия" – распределение оценок испытуемых при выполнении теста), и факторов нестабильности измерительной процедуры (дисперсия ошибок).

[1] П. Клайн не рекомендует использование заданий с ответами типа "истинно-ложно" и добавляет, что при большом количестве заданий влиянием угадывания вообще можно пренебречь.

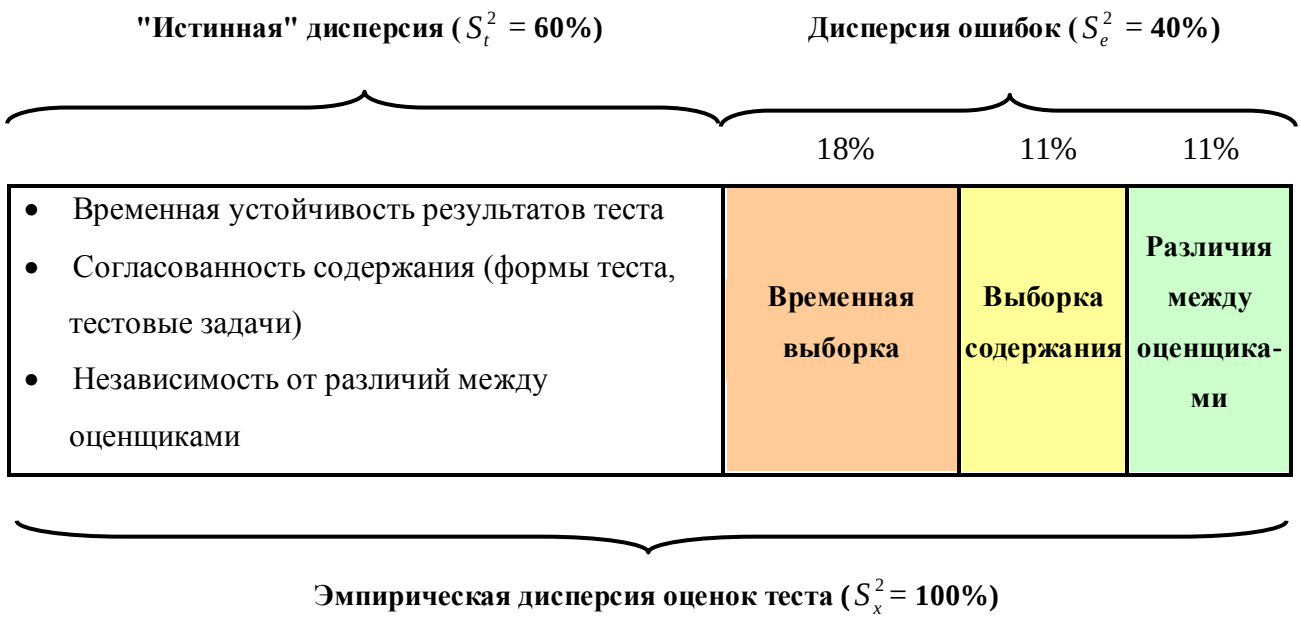


Рисунок 2. Причины, влияющие на дисперсию результатов теста.

- **Классификация основных типов надежности теста**

В самом широком смысле, надежность теста - это характеристика того, в какой степени выявленные у испытуемых различия по тестовым результатам являются отражением действительных ("истинных") различий в измеряемых свойствах и в какой мере они могут быть приписаны действию случайных факторов (ошибкам).

В более узком смысле под этой группой показателей понимают *временную устойчивость результатов теста* - степень согласованности результатов теста, получаемых при первичном и повторном его применении, по отношению к тем же испытуемым в различные моменты времени, с использованием одной формы тесты или разных (но сопоставимых по характеру) наборов тестовых заданий или при других изменениях условий обследования (но при условии, что испытуемые не изменились). Данный тип надежности получил название ***ретестовая надежность*** (*test-retest reliability*). Для ее оценки используются различные коэффициенты корреляции между результатами теста при первичном и повторном применении.

Суть ретестовой надежности заключается в том, что реальные оценки и ранговые места испытуемых при повторном обследовании изменяются, и их распределение в той или иной степени отличается от исходного. При этом дисперсия нового распределения выше исходного на величину дисперсии ошибки измерения. Сказанное можно выразить формулой, описывающей надежность теста (r_t) как отношение «истинной» и реальной

(эмпирической) дисперсии:

$$r_t = \frac{S_t^2}{S_x^2} \text{ или } r_t = 1 - \frac{S_e^2}{S_x^2}$$

где S_t^2 - «истинная» дисперсия (изменчивость измеряемого свойства в группе); S_e^2 - дисперсия ошибки измерения; S_x^2 - эмпирическая дисперсия оценок теста (изменчивость полученных тестовых результатов).

Как видно, надежность теста тесно связана с ошибкой измерения, которая указывает на вероятные пределы колебаний измеряемой величины под воздействием случайных посторонних факторов. Величина ошибки измерения обратно пропорциональна показателям точности измерения. Относительную долю дисперсии ошибки (S_o^2) легко установить, исходя из уравнения:

$$S_o^2 = \frac{S_e^2}{S_x^2} = 1 - r_t$$

Кроме ретестовой надежности в психометрии выделяют также второй тип надежности, связанный с *содержанием инструмента оценки*: ***внутренняя согласованность*** тестовых заданий (*self-consistent*). Для ее измерения не требуется повторного тестирования, а используются различные методы расщепления теста на равноценные части. Коэффициент корреляции между этими частями служит оценкой надежности. Для оценки влияния разбиения теста на его коэффициент надежности используются формулы Спирмена-Брауна, Фланагана, Кристофа. Альтернативным методом вычисления надежности по эквивалентным половинам теста служит метод Рюлона. Кроме анализа частей теста также существуют методы оценки отдельных заданий на их ***однородность*** (*гомогенность*) по содержанию и трудности (*interitem consistency* - «взаимосогласованность заданий»). Для этих целей вычисляются коэффициент Кьюдера-Ричардсона, коэффициент α (альфа) Кронбаха. Данный тип надежности тесно связан с валидностью теста (измеряет ли он то, для измерения чего предназначен.) Если некоторая переменная измеряется частью теста, то тогда в других частях, если они не согласованы с первой, эта же переменная измеряться не может. Из этого следует, что для того, чтобы тест был валидным, он должен быть внутренне согласованным.

Ряд факторов, которые могли бы повлиять на надежность тестовых оценок, хорошо поддаются экспериментальному контролю. Например, ошибки измерения в результате проведения теста в отвлекающей обстановке (шум, посторонние лица) или в более

короткое или длительное, чем это положено, время. Различные формы предъявления инструкций к заданиям. Предъявление двусмысленных, непонятных для обследуемых инструкций и заданий. Время проведения ответственного и напряженного тестирования.

Одним из важнейших средств повышения надежности психодиагностической методики является единообразие процедуры обследования, его строгая регламентация: одинаковые для обследуемой выборки испытуемых время, обстановка и условия работы, однотипный характер инструкций и заданий, одинаковые для всех временные ограничения, способы и особенности контакта с испытуемыми, порядок предъявления заданий и т.д. Обследуемые должны приходить на тестирование отдохнувшими, с хорошим самочувствием. Перед выполнением ответственных и напряженных заданий необходимо проверить самочувствие обследуемых, чтобы оно не влияло на их производительность. При такой стандартизации процедуры исследования можно существенно уменьшить влияние посторонних случайных факторов на результаты теста и таким образом повысить их надежность.

Ниже мы остановимся на тех существенных факторах, влияющих на объективность и эффективность тестовых процедур, контроль которых требует применения специальных психометрических и статистических методов.

ВЛИЯНИЕ ИНСТРУМЕНТА ОЦЕНКИ НА НАДЕЖНОСТЬ ИЗМЕРЕНИЯ

ВРЕМЕННАЯ УСТОЙЧИВОСТЬ РЕЗУЛЬТАТОВ ТЕСТА

- **Ретестовая надежность**

Самый очевидный и понятный метод определения надежности результатов теста - его повторное проведение (***ретестовая надежность*** - *test-retest reliability*). В этом случае коэффициент надежности (r_t) просто равен коэффициенту корреляции между показателями, полученными теми же испытуемыми в каждом из двух случаев проведения теста. Дисперсия ошибок соответствует случайным колебаниям в выполнении заданий от одного сеанса тестирования к другому. Эти колебания могут отчасти быть (см. рис. 3) результатом неконтролируемых условий тестирования, связанных, например, с контекстом оценки (резкие изменения погоды, внезапные шумы и другие отвлекающие факторы) или субъектом оценки (изменения в функциональном состоянии самих тестируемых, вызванные недавними событиями). Ретестовая надежность показывает, в какой степени результаты теста можно распространить на различные случаи его применения. Чем выше надежность, тем менее чувствительны тестовые показатели к случайным суточным изменениям состояния тестируемых и обстановки тестирования.

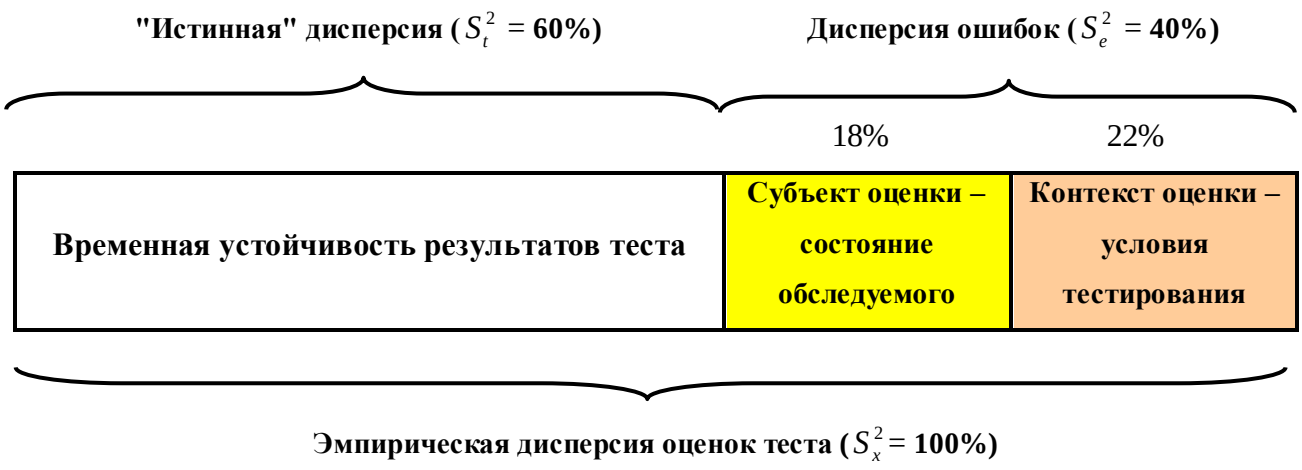


Рисунок 3. Причины, влияющие на результаты теста при повторном тестировании.

Заметим, что использование коэффициента корреляции Пирсона (см. Приложение 1) требует, чтобы результаты тестирования имели нормальное распределение (параметрический критерий). Для проверки предположения о нормальном распределении

выборочных результатов используются критерий Колмогорова-Смирнова или W-критерий Шапиро-Уилка (Shapiro-Wilk). Если гипотеза о нормальности распределения отвергается, можно воспользоваться непараметрическими коэффициентами ранговой корреляции Спирмена (см. Приложение 2) или Кендалла (см. Приложение 3).

При характеристике ретестовой надежности особое значение имеет временной интервал между первым и вторым обследованиями. С его увеличением показатели корреляции имеют тенденцию к снижению, существенно повышается вероятность воздействия посторонних факторов - могут наступить закономерные возрастные изменения измеряемых тестом свойств, произойти различные события, влияющие на состояние и особенности развития исследуемых качеств. По этой причине при определении ретестовой надежности стараются выбирать непродолжительные временные интервалы (до нескольких месяцев), сопровождая сведениями о событиях, происшедших за время между двумя сеансами тестирования с теми, на ком измерялась надежность теста (профессиональная подготовка, работа, семейная жизни т. д.).

П. Клайн полагает, что если между повторными тестированиями прошло много времени, то влияние запоминания испытуемыми своих ответов на результаты теста незначительно, а когда после первого тестирования прошел год, то им можно смело пренебречь. Поэтому он предложил интервал между повторными тестированиями не менее 6 месяцев. Многие авторы полагают, что столь значительного интервала может быть вполне достаточно для того, чтобы произошли изменения в измеряемых поведенческих функциях. Поэтому Анастази утверждает, что интервал между двумя последовательными применениями теста обычно не должен превышать 6 месяцев.

Таким образом, несмотря на кажущуюся простоту и очевидность методики повторного тестирования, ее применение к большинству психологических тестов представляет немалые трудности. Улучшение показателей как результат тренировки при повторении теста будет, вероятно, различным у разных людей. Кроме того, если промежуток времени между первым и вторым тестированием достаточно мал, испытуемые могут припомнить многие из своих прежних ответов. Иными словами, та же картина правильных и ошибочных ответов, вероятно, воспроизводится благодаря работе одной только памяти. Следовательно, результаты двух предъявлений теста не будут независимыми, и корреляция между ними окажется обманчиво высокой. К тому же повторное проведение может изменить саму сущность теста. В первую очередь это относится к задачам, требующим

логических рассуждений или сообразительности. Испытуемый, однажды ухватив принцип решения или построив всю цепь рассуждений, в дальнейшем может воспроизводить правильный ответ, минуя промежуточные ступени. Методика повторного тестирования применима только к тем тестам, на которые их повторное проведение на одних и тех же испытуемых не оказывает заметного влияния. К этой категории относится ряд моторных тестов и тестов сенсорного различения. Однако для подавляющего большинства психологических тестов эта методика определения коэффициента надежности оказывается малоэффективной.

- **Надежность параллельных форм теста**

Один из способов нивелировать трудности, с которыми приходится сталкиваться при определении ретестовой надежности, это использование взаимозаменяемых форм (*alternate forms*) теста. Одних и тех же испытуемых могут тестировать в первый раз с помощью одной формы, а второй раз - с помощью другой, эквивалентной формы. Корреляция между показателями, полученными по двум формам теста, представляет его коэффициент надежности (***надежность параллельных или взаимозаменяемых форм***). Заметим, что такой коэффициент надежности служит мерой как временной устойчивости результатов теста, так и согласованности ответов на различные формы теста (см. рис. 4). Таким образом, этот коэффициент служит смешанной характеристикой двух типов надежности.

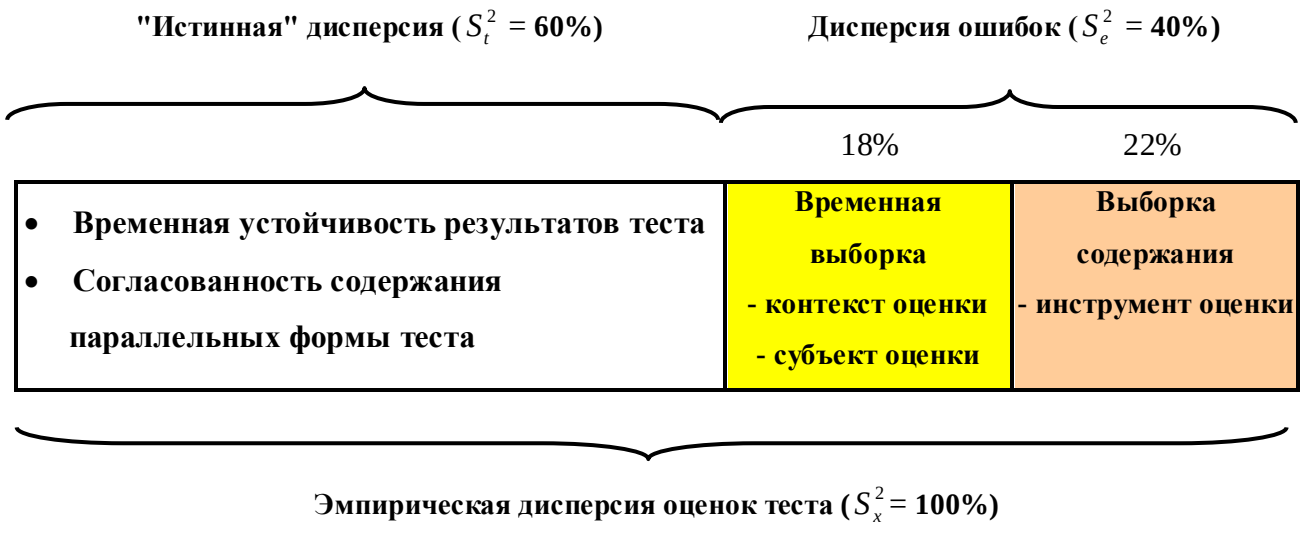


Рисунок. 4. Причины, влияющие на результаты теста при использовании параллельных форм.

Как и в случае ретестовой надежности, сведения о надежности взаимозаменяемых форм всегда должны сопровождаться указанием длительности временного интервала между двумя предъявлениями теста, а также характеристикой релевантных событий, происшедших за это время в жизни испытуемых. Если обе формы применяются непосредственно одна за другой, то полученная корреляция характеризует только согласованность содержания параллельных форм теста (отражает их взаимозаменяемость), но ничего не говорит о надежности как временной устойчивости. Дисперсия ошибок в этом случае обусловлена колебаниями результатов при переходе от одного набора заданий к другому, а не временными флуктуациями показателей (см. рис. 5).

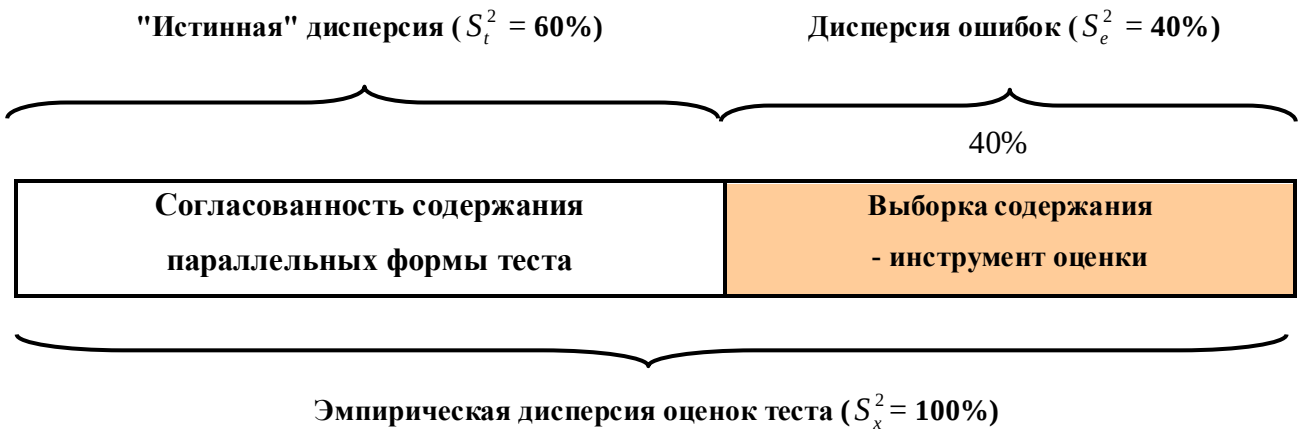


Рисунок. 5. Причины, влияющие на результаты теста при непосредственном использовании параллельных форм.

При разработке взаимозаменяемых форм, безусловно, следует позаботиться о том, чтобы они на самом деле были параллельными. Принципиально важно, чтобы параллельные формы конструировались как независимые тесты, отвечающие, однако, одним и тем же требованиям. Такие тесты должны содержать одинаковое число заданий, представленных в одной и той же форме и с однотипным содержанием. Диапазон и уровень трудности заданий тоже должны быть одинаковыми. Инструкции, временные рамки, поясняющие примеры, формат бланков и все другие аспекты теста также необходимо проверить на сопоставимость.

Следует добавить, что наличие параллельных форм желательно и по другим соображениям, помимо определения надежности теста. Взаимозаменяемые формы полезны при повторных исследованиях и при изучении влияния некоторых промежуточных экспериментальных факторов на выполнение теста. Использование нескольких взаимозаменяемых форм

служит, кроме того, средством уменьшения возможности натаскивания в выполнении тестов и обмана.

Несмотря на гораздо более широкое, сравнительно с ретестовой надежностью, применение, надежность взаимозаменяемых форм также обнаруживает ряд ограничений. Прежде всего, если изучаемые поведенческие функции подвержены значительному влиянию тренировки, использование параллельных форм ослабит, но не устранит его полностью. Конечно, если бы у всех тестируемых наблюдалось одно и то же улучшение результатов при повторном проведении теста, это не повлияло бы на корреляцию показателей, поскольку прибавление постоянной величины к каждому показателю не меняет коэффициента корреляции. Однако, скорее всего, улучшение результатов у разных людей будет неодинаковым вследствие индивидуальных различий в опыте работы с подобным материалом, в мотивации участия в тесте и по другим причинам. При этих условиях эффект тренировки представляет собой еще один источник дисперсии, снижающей, в общем, корреляцию между двумя формами. Но если влияние тренированности невелико, снижение корреляции будет незначительным.

Другая проблема связана с возможным изменением сущности теста при повторном его проведении. Например, если в параллельных задачах на сообразительность применен один и тот же принцип, то большинство испытуемых, однажды найдя решение, и во второй раз применят его. В подобных случаях одной замены содержания заданий явно недостаточно для того, чтобы избежать переноса принципа принципов решения из одной формы теста на другую. Наконец, следует добавить, что для многих тестов взаимозаменяемые формы отсутствуют ввиду практических трудностей создания подлинно эквивалентных форм. В силу этих причин часто приходится обращаться к другим методам оценки надежности теста.

- **Коэффициент ассоциации Пирсона**

Л.Ф. Бурлачук и С.М. Морозов [6] предлагают при расчете ретестовой надежности и надежности параллельных форм кроме традиционных коэффициентов корреляции использовать φ (фи) коэффициент ассоциации Пирсона для номинальных дихотомических данных (переменные измерены по дихотомической шкале, например: соответствие или несоответствие "ключу" ответа на вопрос):

$$\varphi = \frac{p_{xy} - p_x \cdot p_y}{\sqrt{p_x \cdot q_x \cdot p_y \cdot q_y}}$$

где p_x, p_y – доля случаев соответствия ответа на вопрос "ключу" по переменным X и Y; q_x и q_y – доля случаев несоответствия ответа на вопрос "ключу" по X и Y; $q = 1 - p$; p_{xy} – доля случаев соответствия ответа на вопрос "ключу" как по X, так и по Y.

Пример вычисления φ - коэффициента ассоциации Пирсона при сравнении параллельных форм опросника представлен в таблице 1. Используя критерий χ^2 (при степени свободы $df = 1$), можно определить значимость связи, установленной φ - коэффициентом. В нашем примере $\chi^2 = n \times \varphi^2 = 12 \times (-0.44)^2 = 0.145$; $\chi^2 < \chi^2_{кр}$ ($p > 0.1$), связь несущественна.

Номер вопроса	**Форма "A" (X)**	**Форма "B" (Y)**	**Вычисление**
1	0	0	$p_x = 0.583$; $q_x = 0.417$
2	1	1	$p_y = 0.333$; $q_y = 0.667$
3	0	1	$p_{xy} = 0.167$
4	0	0	
5	1	0	$\varphi = \frac{0.167 - 0.583 \cdot 0.333}{\sqrt{0.583 \cdot 0.417 \cdot 0.333 \cdot 0.667}}$
6	1	0	
7	0	1	$\varphi = -0.44$
8	1	1	
9	0	0	
10	1	0	
11	1	0	
12	1	0	
Примечание: 0 – несовпадение ответа с "ключом"; 1 – совпадение ответа с "ключом".			

Таблица 1. Вычисление φ - коэффициента ассоциации Пирсона при сравнении параллельных форм опросника.

- **Коэффициент корреляции Гилфорда и каппа-коэффициент**

Другой подход вычисления φ - коэффициента для оценки надежности параллельных форм основывается на использовании матрицы сопряженности результатов критериально-ориентированного теста[2]. Пусть в зависимости от числа правильных ответов обучаемый получает либо "зачет", либо "незачет" (критерий устанавливается экспертами). Таблица 2 представляет собой пример матрицы сопряженности 2 × 2, где *a*, *b*, *c* и *d* – доли испытуемых, получивших соответствующую аттестацию по результатам двух параллельных форм теста.

		Параллельная форма теста *А*	
		Зачет	**Незачет**
Параллельная форма теста *Б*	**Зачет**	*a* = 0.50	*b* = 0.00
	Незачет	*c* = 0.05	*d* = 0.45

Таблица 2. Результаты тестирования параллельных форм теста.

На основе матрицы сопряженности вычисляются следующие коэффициенты надежности:

- ***φ - коэффициент корреляции Гилфорда***

$$\varphi = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

- ***каппа-коэффициент*** (**k**)

$$k = \frac{P - P_c}{1 - P_c}$$

где $P = a + d$ – вероятность принятия согласованного решения по матрице сопряженности результатов критериально-ориентированного теста, представляющая собой сумму вероятностей принятия согласованных решений по отнесению испытуемых в каждую из групп; $P_c = (c + d)\times(d + b) + (a + b)\times(c + a)$ – вероятность случайного согласования.

[2] ***Критериально-ориентированный тест*** – система заданий, позволяющая измерить уровень индивидуальных достижений относительно полного объема знаний, навыков и умений, которые должны быть, например, усвоены в процессе подготовки (устанавливается критерий по числу правильных ответов, который позволяет судить об усвоение конкретной программы). ***Нормативно-ориентированный тест*** – позволяет сравнивать достижения в деятельности, обучении (уровень профессиональных знаний, умений, навыков) отдельных обследуемых друг с другом (общее число правильных ответов).

Каппа-коэффициент надежности теста представляет собой оценку надежности критериально-ориентированного теста, учитывающую случайную согласованность. Он изменяется в диапазоне [–1; +1].

Значимость φ - коэффициента корреляции Гилфорда определяется его сравнением с критическим значением коэффициента:

$$\varphi_{êðèò} = \sqrt{\frac{\chi^2}{a+b+c+d}}$$

где χ^2 – стандартный квантиль распределения "хи-квадрата" с заданным уровнем значимости и одной степенью свободы df (для $p = 0.05$ $\chi^2 = 3.84$, для $p = 0.01$ $\chi^2 = 6.62$).

Если модуль φ - коэффициента корреляции Гилфорда не меньше критического значения, то связь результатов выполнения параллельных форм критериально-ориентированного теста считается установленной (с заданным уровнем значимости) и можно сделать вывод о надежности исследуемого теста.

Говорить о надежность проведенных тестовых испытаний параллельных форм критериально-ориентированного теста можно по равенству φ и каппа коэффициентов надежности, а также при превышении ими значений 0.8.

- **Тетрахорический коэффициент корреляции Пирсона**

Если две параллельные формы теста эквивалентны по уровню трудности, а уровень, например, подготовки обучаемых описывается нормальным распределением баллов в параллельных формах теста, то можно рассчитать надежность по так называемому ***тетрахорическому коэффициенту корреляции Пирсона***:

$$r_{tet} = \cos\left(\frac{\pi}{1+\sqrt{\frac{ad}{bc}}}\right)$$

Этот коэффициент независим от среднего уровня способностей испытуемых, но зависит от уровня индивидуальных способностей и будет выше для более разнородно подготовленной группы тестируемых. Если распределение баллов в параллельных тестах не соответствует

нормальному, то тетрахорический коэффициент корреляции Пирсона дает завышенную оценку и не может быть применим.

СОГЛАСОВАННОСТЬ (ОДНОРОДНОСТЬ) СОДЕРЖАНИЯ ИНСТРУМЕНТА ОЦЕНКИ

- **Метод расщепления**

Меру надежности можно определить и на основании однократного применения единственной формы теста, пользуясь для этого различными процедурами расщепления теста на две равноценные половины. При таком способе каждый испытуемый получает два показателя благодаря разделению теста на две эквивалентные части. Очевидно, что надежность, найденная методом расщепления, дает нам меру согласованности выборок содержания. Временная устойчивость показателей в такой характеристике надежности не представлена, поскольку она предполагает только один сеанс тестирования (см. рис. 6). Этот тип коэффициента надежности иногда называют ***коэффициентом внутренней согласованности***, так как для его определения требуется лишь однократное проведение единственной формы теста.

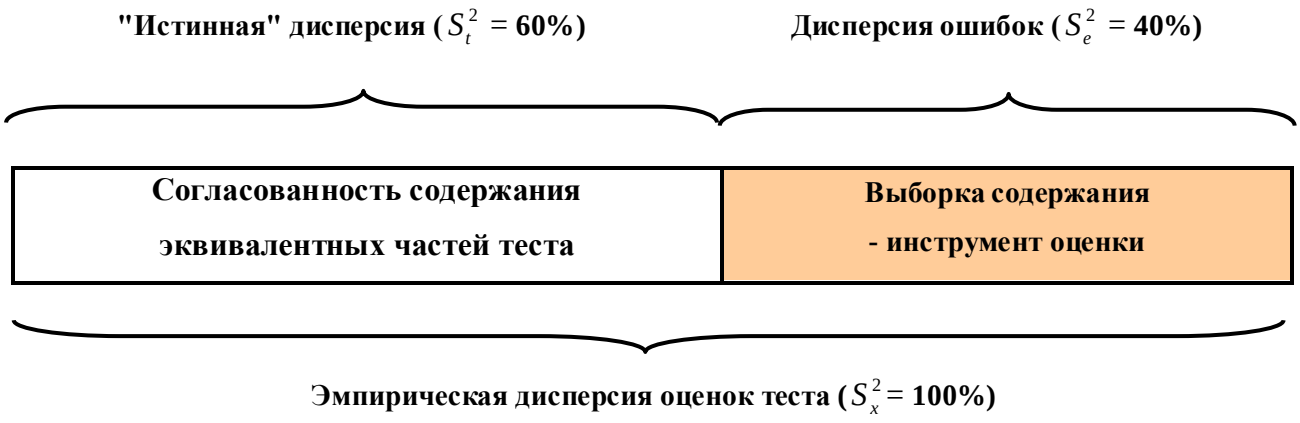


Рисунок 6. Причины, влияющие на результаты теста - метод расщепления.

Первая проблема, с которой мы сталкиваемся при применении метода расщепления, связана с тем, как разделить тест, чтобы добиться максимальной эквивалентности его половин. Всякий тест можно членить многими способами. В большинстве тестов первая и вторая половины оказались бы неэквивалентными вследствие различий в характере и уровне трудности заданий, а также в связи с кумулятивными эффектами вхождения в работу, практики, утомления, скуки и любых других факторов, воздействие которых

нарастает от начала к концу теста. Подходящий для большинства целей метод состоит в вычислении показателей отдельно по четным и нечетным заданиям теста. Если задания теста были изначально расположены в порядке возрастания трудности, то такое разбиение дает практически эквивалентные показатели обеих половин. Одна предосторожность, которую требуется при этом соблюдать, относится к случаю, когда тест содержит группу взаимосвязанных заданий - например, когда несколько вопросов касаются какого-то одного чертежа механического устройства в тесте технических способностей или одного и того же фрагмента текста в тесте чтения. В этом случае каждая такая группа заданий должна быть целиком отнесена либо к одной, либо к другой половине. Если задания таких групп разделить на две части, то возникнет обманчивое сходство сравниваемых показателей, так как любая ошибка в понимании задачи скажется на выполнении заданий из обеих половин.

Полученные показатели по двум частям теста коррелируются обычным методом. Нужно иметь в виду, однако, что эта корреляция в действительности показывает надежность лишь половины теста. Например, если весь тест состоит из 100 заданий, то корреляция вычисляется между двумя множествами показателей, каждый из которых основан только на выполнении 50 заданий. В отличие от надежности этого типа, при расчете ретестовой надежности, как и надежности взаимозаменяемых форм, каждый показатель основывается на полном наборе заданий теста.

- **Формула Спирмена-Брауна**

При прочих равных условиях, чем больше заданий содержит тест, тем выше его надежность. Вполне оправданно ожидать, что чем обширнее выборка поведения, тем адекватнее и согласованнее получаемые единицы измерения. Влияние, оказываемое увеличением или сокращением теста на его коэффициент надежности, можно оценить с помощью ***формулы Спирмена-Брауна***:

$$r_{tt} = \frac{n \times r_{hh}}{1 + (n-1) \times r_{hh}}$$

где r_{tt} - ожидаемое значение коэффициента надежности; r_{hh} - полученное значение коэффициента надежности; n - отношение нового числа заданий к первоначальному. Так, если число заданий теста возросло с 25 до 100, то $n = 4$, а если оно сократилось с 60 до 30, то $n = 0.5$.

Формула Спирмена-Брауна широко используется при определении надежности методом расщепления, и во многих руководствах к тестам данные о надежности приводятся в этом

виде. Применительно к расчетам надежности эквивалентных частей теста формула Спирмена-Брауна всегда предполагает удвоение числа заданий теста, и потому может быть приведена к более простому виду:

$$r_{tt} = \frac{2 \cdot r_{hh}}{1 + r_{hh}},$$

где r_{hh} - корреляция эквивалентных половин теста.

Допустим, коэффициент надежности половины теста (найденный методом расщепления теста на две эквивалентные части) оказался равным 0.78, тогда ожидаемое значение коэффициента надежности по всему тесту будет равно:

$$r_{tt} = \frac{2 \cdot 0.78}{1 + 0.78} = 0.88$$

- **Формула Фланагана**

Приведенные формулы справедливы для случаев равных стандартных отклонений обеих половин теста ($\sigma_{x_1} = \sigma_{x_2}$). Если σ_{x_1} отличается от σ_{x_2}, то для определения коэффициента надежности применяется ***формула Фланагана***:

$$r_{tt} = \frac{4 \cdot \sigma_{x_1} \cdot \sigma_{x_2} \cdot r_{hh}}{\sigma_{x_1}^2 + \sigma_{x_2}^2 + 2 \cdot \sigma_{x_1} \cdot \sigma_{x_2} \cdot r_{hh}}$$

- **Формула Кристофа**

Этот же показатель для малых выборок рассчитывается по ***формуле Кристофа***:

$$r_{tt} = \frac{2}{n-1} + \frac{n-3}{n-1} \times \frac{4 \cdot \sigma_{x_1} \cdot \sigma_{x_2} \cdot r_{hh}}{\sigma_{x_1}^2 + \sigma_{x_2}^2 + 2 \cdot \sigma_{x_1} \cdot \sigma_{x_2} \cdot r_{hh}}$$

- **Формула Рюлона**

Альтернативный метод вычисления надежности эквивалентных половин теста был разработан Рюлоном. Требуется знать только дисперсию разностей между показателями каждого испытуемого по обеим половинам теста (σ_Δ^2 или S^2_d) и дисперсию показателей по полному тесту (σ_x^2 или S^2_x); значения этих величин подставляются в следующую ***формулу Рюлона***, которая позволяет сразу получить характеристику надежности полного теста:

$$r_t = 1 - \frac{\sigma_\Delta^2}{1 + \sigma_x^2}$$

Интересно отметить связь между этой формулой и определением дисперсии ошибок. Любая разность между показателями испытуемого по двум половинам теста отражает постороннее влияние или дисперсию ошибок. Дисперсия таких разностей, поделенная на дисперсию показателей по всему тесту, дает долю дисперсии ошибок в этих показателях. Вычитая эту дисперсию ошибок из единицы, мы получаем долю «истинной» дисперсии для установленного применения теста, которая равна его коэффициенту надежности.

- **Формула Спирмена-Брауна**

Надежность теста связана со средней корреляцией между заданиями, то есть с его однородностью. Для этих целей используют ***формулу Спирмена-Брауна*** в следующем виде:

$$r_{tt} = \frac{n \times \bar{r}_{ij}}{1 + (n-1) \times \bar{r}_{ij}}$$

где r_{tt} - ожидаемое значение коэффициента надежности; $\bar{r}_{ij}$ средняя взаимная корреляция между заданиями; n – количество заданий.

Пусть у нас есть три набора заданий (10, 20, 30) и средняя корреляция между заданиями равна 0.20 (достаточно низкая взаимная корреляция). В этом случае мы получим следующие значения надежности теста (см. табл. 3):

Число заданий теста	**Расчет коэффициента надежности**
10	$r_{tt} = \frac{10 \times 0.20}{1 + (9 \times 0.20)} = 0.667$
20	$r_{tt} = \frac{20 \times 0.20}{1 + (19 \times 0.20)} = 0.800$
30	$r_{tt} = \frac{30 \times 0.20}{1 + (29 \times 0.20)} = 0.959$

Таблица 3. Надежность теста при различном числе заданий.

Если мы подставим в формулу более высокую среднюю корреляцию между заданиями (0.40), то для набора из 30 заданий надежность теста, рассчитанная по взаимным корреляциям заданий, будет равна:

$$r_{tt} = \frac{30 \times 0.40}{1 + (29 \times 0.40)} = 0.923$$

Следует заметить, что если разработчик смог создать тест из 30 однородных заданий, то расщепление теста на две параллельные форма по 15 заданий также будет иметь удовлетворительную надежность.

Использование приведенной выше формулы Спирмена-Брауна для вычисления надежности теста по оценки однородности его задач требует трудоемких вычислений корреляционных матриц. Поэтому были разработаны другие методы: метод Кьюдера-Ричардсона, коэффициент α Кронбаха.

- **Метод Кьюдера-Ричардсона**

Данный метод также использует однократное предъявление единственной формы теста, но при этом оцениваются отдельные задания на их однородность (гомогенность) по содержанию и трудности. При гетерогенных заданиях показатель надежности будет снижаться. Самая распространенная методика оценки гомогенности заданий дихотомического типа была разработана Кьюдером и Ричардсоном (Kuder, Richardson). Как и в методах расщепления, внутренняя согласованность находится по данным однократного проведения единственной формы теста, но вместо использования показателей по двум эквивалентным половинам теста эта методика опирается на результаты выполнения каждого задания (см. рис. 7).

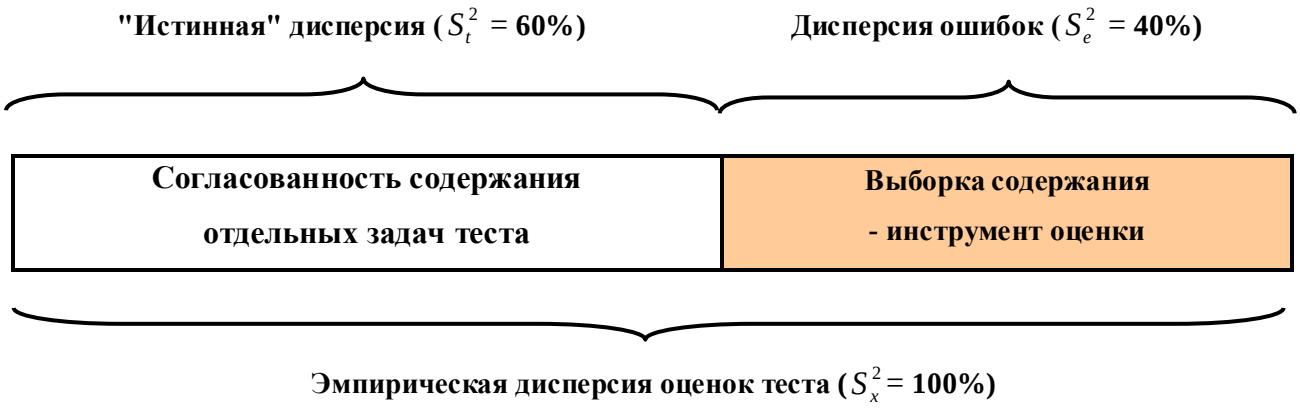


Рисунок 7. Причины, влияющие на результаты теста – оценка гомогенности заданий.

Наиболее распространенным методом оценки надежности отдельных заданий является вычисление ***коэффициента Кьюдера-Ричардсона:***

$$r_{tt} = \frac{SD_t^2 - \sum pq}{2 \cdot SD_t^2} + \sqrt{\frac{\sum r_{pb}^2 pq}{SD_t^2} + \left(\frac{\sigma_x^2 - \sum pq}{2SD_t^2} \right)}$$

где r_{tt} - коэффициент надежности полного теста; r_{pb} - коэффициент дискриминации (см. Коэффициент дискриминации); SD_t - стандартное отклонение суммарных показателей теста (оценок по тесту); p и q - доля испытуемых, соответственно справившихся (p) и не справившихся (q) с каждым заданием. Чтобы вычислить $\sum pq$, нужно для каждого задания найти произведение $p \times q$, а затем сложить эти произведения по всем заданиям.

- **Формула Гуликсена**

В целях упрощения вычисления может быть применена **формула Гуликсена**:

$$r_{tt} = \frac{k}{k-1} \cdot \left[1 - \frac{\sum pq}{(\sum r_{pb} \cdot \sqrt{pq})^2} \right]$$

где r_{tt} - коэффициент надежности полного теста; r_{pb} - коэффициент дискриминации (см. Коэффициент дискриминации); p и q - доля испытуемых, соответственно справившихся (p) и не справившихся (q) с каждым заданием. Чтобы вычислить $\sum pq$, нужно для каждого задания найти произведение $p \times q$, а затем сложить эти произведения по всем заданиям; k - число задач в тесте.

- **Формула KR-20**

При отсутствии коэффициента дискриминации (r_{pb}) применим следующий часто используемый вариант формулы Кьюдера-Ричардсона (***KR-20***):

$$r_{tt} = \left(\frac{k}{k-1} \right) \cdot \left(1 - \frac{\sum pq}{SD_t^2} \right)$$

где r_{tt} - коэффициент надежности полного теста; SD_t - стандартное отклонение суммарных показателей теста (оценок по тесту); p и q - доля испытуемых, соответственно справившихся (p) и не справившихся (q) с каждым заданием. Чтобы вычислить $\sum pq$, нужно для каждого задания найти произведение $p \times q$, а затем сложить эти произведения по всем заданиям; k - число задач в тесте.

В таблице 4 приведен пример вычисления r_{tt} по методу Кьюдера-Ричардсона.

Можно математически доказать, что коэффициент надежности Кьюдера-Ричардсона представляет собой среднее значение коэффициентов надежности частей теста (формула Рюлона), вычисляемых для всех возможных разбиений теста надвое. Обычный же коэффициент надежности частей теста основан на разбиении, построенном в расчете на получение эквивалентных половин. Поэтому в случае неоднородности заданий теста

коэффициент Кьюдера-Ричардсона будет ниже коэффициента надежности эквивалентных половин. Фактически, разность между этими двумя коэффициентами может служить приблизительной числовой оценкой однородности теста.

Номер задачи	**Количество лиц, решивших задачу (N^+)**	***p*** **(N/N^+)**	***q*** **($1-N/N^+$)**	***pq***
1	48	0.96	0.04	0.04
2	43	0.86	0.14	0.12
3	33	0.66	0.34	0.22
4	39	0.78	0.22	0.17
5	28	0.56	0.44	0.25
…				
15	1	0.02	0.98	0.02
16	1	0.02	0.98	0.02
				Σpq = 2.55
$r_{tt} = \left(\frac{k}{k-1}\right)\cdot\left(1-\frac{\sum pq}{SD_t^2}\right) = \frac{16}{15}\cdot\left(1-\frac{2.55}{8.01}\right) = 0.72$				

Таблица 4. Пример определения коэффициента надежности методом Кьюдера – Ричардсона. (SD^2_t = 8.01; k = 16; число обследованных N = 50)

Если величина коэффициента надежности *KR-20* составляет от 0.90 до 0.99, то тест имеет отличную оценку надежности, если от 0.80 до 0.89 то хорошую, от 0.70 до 0.79 – удовлетворительную и менее 0.69 - неудовлетворительную надежность.

- **Коэффициент Кронбаха**

Формула Кьюдера-Ричардсона применима лишь к тем тестам, в которых выполнение заданий оценивается в дихотомической шкале, по принципу «выполнено – не выполнено». Для случаев с более дифференцированной оценкой (например, по 5-балльной шкале) применима формула «***коэффициента α Кронбаха***», которая считается наиболее эффективной для оценки надежности. В этой формуле $\sum pq$ заменена на $\sum(SD^2_i)$ - сумму дисперсий балльных оценок по каждому заданию теста. Процедура вычислений состоит в

нахождении дисперсии всех индивидуальных балльных оценок по каждому заданию с последующим суммированием этих дисперсий по всем заданиям. Полная ***формула коэффициента α Кронбаха*** выглядит следующим образом:

$$\alpha = \left(\frac{k}{k-1}\right) \cdot \left(1 - \frac{\sum(SD_i^2)}{SD_t^2}\right)$$

где SD_t - стандартное отклонение суммарных показателей теста (оценок по тесту); $\sum(SD^2_i)$ - сумму дисперсий балльных оценок по каждому заданию теста; k - число задач в тесте.

В общем виде, коэффициент α (альфа) определяется как оценка корреляции данного теста с другим тестом такой же длины из одной генеральной совокупности заданий. Из формулы коэффициента α можно сделать вывод, что надежные тесты имеют большую дисперсию балльных оценок по каждому заданию теста (и, следовательно, являются более дискриминативными), чем ненадежные тесты. Коэффициент α Кронбаха является одним из самых часто используемых при оценке внутренней согласованности тестовых заданий.

- **Формула Спирмена-Брауна для оценки надежность теста с изменением его длины**

Используя коэффициента α Кронбаха можно оценить надежность теста с изменением его длины. Для оценки надежности более длинного (или более короткого) теста, при известном значении коэффициента α Кронбаха, используют следующую формулу Спирмена-Брауна:

$$\alpha_{new} = \frac{m \cdot \alpha_{old}}{1 + (m-1) \cdot \alpha_{old}}$$

где α_{new} - это новая оценка надежности после удлинение (или укорачивания) теста, α_{old} - оценка надежности теста с начальной длиной, m = отношению новой длины теста к начальной длине теста. Важно отметить, что для корректного использования формулы Спирмена-Брауна необходимо, чтобы задания, добавляемые для увеличения теста, должны быть такого же качества, как и первоначальные задания теста.

Характеристика надежности по типу надежности частей теста или отдельных задач имеет серьезные преимущества по сравнению с надежностью ретестовой и надежностью параллельных форм, главным образом благодаря отсутствию необходимости в повторном обследовании. Таким образом, снимается влияние многих посторонних факторов, в частности тренировки, запоминания решений и т. д. Это обстоятельство определяет широкое распространение методов характеристики надежности частей теста или отдельных

задач по сравнению с другими типами надежности. К недостаткам метода относят невозможность установить устойчивость результатов теста спустя определенное время (временная устойчивость). Это требует комбинирования метода надежности частей теста или отдельных задач с другими типами характеристики надежности психодиагностической методики.

- **Метод дисперсионного анализа**

Данный метод был предложен Хойтом, который рассматривал ответы на задания как двухфакторный анализ дисперсий без репликации (повторения). Данный метод идентичен применению коэффициента α и формуле K-R20, и может служить альтернативой для оценки однородности заданий:

$$r_{tt} = 1 - \frac{V_r}{V_e}$$

где V_r — дисперсия остатка от суммы квадратов, а V_e — дисперсия для испытуемых.

$$V_e = \frac{\sum d_e^2}{N-1},$$

где $\sum d_e^2$ - сумма квадратов для испытуемых:

$$\sum d_e^2 = \frac{\sum X_t^2}{n} - \frac{\left(\sum X_t\right)^2}{n \cdot N}$$

где X_t — общий показатель по тесту для каждого испытуемого, n — количество заданий теста, N — количество испытуемых.

$$V_r = \frac{1}{N \cdot n - N - n + 1} \cdot \left(\frac{\sum R_i \cdot \sum W_i}{\sum R_i + \sum W_i} - \sum d_e^2 - \sum d_i^2 \right),$$

где R_i — количество правильных ответов для задания i; W_i — количество неправильных ответов для задание i ($W_i = N - R_i$); $\sum d_i^2$ - сумма квадратов для заданий:

$$\sum d_i^2 = \frac{\sum R_i^2}{N} - \frac{\left(\sum X_t\right)^2}{n \cdot N}.$$

СОГЛАСОВАННОСТЬ В СКОРОСТИ ВЫПОЛНЕНИЯ ТЕСТОВЫХ ЗАДАНИЙ

Когда время выполнения теста лимитировано, целесообразно исследовать согласованность в скорости выполнения тестовых заданий. Для определения надежности тестов с выраженным скоростным компонентом применяют метод повторного тестирования («тест -

ретест»), метод взаимозаменяемых, эквивалентных форм. Можно воспользоваться и методом расщепления при условии, что задания теста разбиваются по временным характеристикам, а не по порядковым номерам, как при расчете коэффициентов надежности половин теста и Кьюдера-Ричардсона (оба метода основаны на учете согласованности числа ошибок, сделанных испытуемым). Иными словами, показатели по половинам теста должны основываться на раздельно нормированных по времени частях теста. Одним из способов такого разделения является проведение двух эквивалентных половин теста с отдельно устанавливаемыми временными пределами. Например, четные и нечетные задания распечатываются на разных листах и по каждому набору заданий устанавливается временной лимит, равный половине лимита для всего теста. Такая процедура равносильна проведению следующих друг за другом двух эквивалентных форм теста. Хотя каждая форма вдвое короче целого теста, показатели тестируемых, как обычно, основываются на результатах выполнения всего теста. По этой причине, чтобы определить надежность полного теста, нужно воспользоваться соответствующими формулами Спирмена-Брауна, Фланагана или Кристофа.

Если раздельное проведение двух половин теста невозможно, то вместо этого можно воспользоваться разделением полного времени теста на четыре части с регистрацией результатов отдельно для каждой четверти. Это легко осуществить, прося испытуемых по условленному сигналу проводящего тест отметить крестиком выполняемое в данный момент задание (ситуации еще более упрощается при использовании компьютеризированных программ). Число заданий, правильно выполненных за первую и четвертую части полного временного лимита, можно затем объединить для вычисления показателя по первой половине теста. Показатель по другой половине теста будет тогда соответствовать числу заданий, с которыми испытуемый справился за вторую и третью четверти. Такая комбинация четвертей способствует нейтрализации кумулятивных эффектов тренировки, утомления и других факторов. Этот метод особенно хорошо работает, когда задания не отличаются резко друг от друга по уровню трудности.

НЕЗАВИСИМОСТЬ ОЦЕНОК ОТ РАЗЛИЧИЙ МЕЖДУ ОЦЕНЩИКАМИ

Большинство тестов, особенно если они предназначены для массового обследования с использованием компьютеров для вычисления показателей, настолько стандартизированы, что их проведение и регистрация результатов сводят на нет дисперсию ошибок,

обусловленную этими факторами. Пользуясь такими тестами, необходимо лишь внимательно следить за выполнением соответствующих предписаний. Вместе с тем, например, при использовании оценочных шкал для характеристики деятельности индивидуума, возможна значительная дисперсия оценок наблюдателей (scorer variance). При работе с такими методами потребность в мере надежности оценщика столь же велика, как и в более привычных коэффициентах надежности.

- **Ошибки оценщика (эксперта)**

Ошибками оценщика *(rater errors)* называют искажения, допускаемые наблюдателями при использовании оценочных шкал для характеристики деятельности индивидуума. Уровень компетентности оценщика, как и его пол, социальный статус, возраст, - оказывают влияние на выносимые им суждения. Хотя большинство таких ошибок имеют специфический характер, существуют распространенные типы ошибок оценщиков (экспертов), проявляющиеся в широком спектре ситуаций.

Ошибка снисходительности (позитивного уклона – "мягкая" оценка, *leniency error*) возникает, когда средние оценки имеют тенденцию превышать среднюю точку шкалы вследствие: а) давления на оценщика необходимой высокой оценки подчиненных; б) ощущения того, что оценка подчиненного отражает оценку самого оценщика; в) предварительного отбора учащихся или подчиненных перед процедурой оценивания. Эта ошибка приводит к стиранию различий между оцениваемыми людьми. ***Ошибка суровости*** (негативного уклона – "строгая" оценка) является оборотной стороной той же самой ситуации. Если в первом случае оценки группируются в верхней части шкалы, то во втором – в нижней.

Ошибка центральной тенденции (*error of central tendency*) возникает, когда оценщик постоянно выбирает среднюю область значений шкалы, избегая крайних участков. Это может происходить вследствие колебаний в своем праве «быть Господом Богом» или потому, что крайние оценки (неудовлетворительные или плохие) требуют дополнительной поддержки и могут серьезно сказаться на последующих взаимоотношениях оценщика с его подчиненными.

Эффект ореола или ***эффект дьявола*** (***гало-эффект***) возникает, когда одна личностная черта оказывает влияние на характер измерения всех остальных. Эффект ореола выражается в положительной генерализации на другие черты, эффект дьявола выражается

в отрицательной генерализации. Например, работник, получивший высокую оценку по качеству работы, будет также высоко оценен по количеству, инициативе, кооперации и т.д. Причиной гало-эффекта является ряд факторов: личные отношения и чувства, первоначальные ожидания и др.

Ошибка последовательности появляется в тех случаях, когда специфический порядок оцениваемых черт оказывает специфическое воздействие на оценку последующих черт, такое как эффект ореола.

Логическая ошибка возникает, когда оценщик коррелирует специфические черты на основе их предполагаемой согласованности (по эффекту) с другими чертами. Логическая ошибка носит более сложный характер, чем ошибка ореола.

Эффект недавности (или ***эффект новизны***) возникает, когда случай, произошедший незадолго до процедуры оценки, оказывает на оценщика большее влияние, нежели это имело бы место, произойди он гораздо раньше. Особенную проблему здесь представляют события эмоционального характера, например, трудовой конфликт, несчастный случай, ссора. Также надо помнить, что на оценку в большей степени влияют последние события и наблюдения, вплотную предшествующие оценочной сессии и ясно запечатленные в памяти оценщика. Как правило, перед формальной процедурой оценки люди особенно старательно относятся к работе, чтобы показать себя с лучшей стороны. Если оценщик для принятия решения по оценке будет рассматривать только этот период, то вероятность ошибки “новизны” будет очень высокой.

Эффект контрастности возникает в том случае, когда оценка деятельности одного работника влияет на оценку другого. Так, работник, чья деятельность заслуживает средней оценки, оцениваемый сразу же вслед за кем-либо, получившим низкую оценку, получает более высокий рейтинг. Может быть и в точности наоборот, если этот работник “имеет неудачу” быть оцениваемым сразу после высоко оцененного сослуживца. В данной ситуации имеет место эффект “контраста”.

- **Основные меры по снижению ошибок оценщика (эксперта)**

Следующие шаги, по мнению специалистов, помогают предотвратить или, по крайней мере, снизить ошибки в оценке:

1. Для предупреждения ошибок при оценивании шкалу оценки рекомендуется конструировать таким образом:

- критерии, включенные в шкалу, должны быть значимыми для характеристики деятельности, ясно сформулированными и четко определены в точных и конкретных терминах;
- каждому измерению должен соответствовать строго один критерий деятельности работников;
- субъективные оценки должны выражаться в форме, которая бы одинаково интерпретировалась всеми оценщиками;
- точки градации на шкале рейтинга (“отлично”, “хорошо”, ... и т. д.) должны быть недвусмысленно определены в терминах профессионального поведения работников; вместо использования чисел или общих описательных характеристик, в которые различные оценщики вкладывают разный смысл, степень выраженности той или иной черты лучше определять через тщательно сформулированные поведенческие эталоны.

2. Для исключения эффекта “контраста” следует избегать проведения оценки большого количества людей в малый промежуток времени.

3. Оценщики должны пройти соответствующее обучение для ознакомления со всеми разновидностями ошибок в целях недопущения их в оценочной практике.

Также для устранения влияния ошибок оценщика предлагается использовать:

- поведенчески выверенные оценочные шкалы,
- ранжирование персонала по выбранным критериям (сохраняются ошибки эффекта "ореола" и предвзятости),
- метод парных сравнений,
- метод вынужденного выбора и принудительного распределения.

Остановимся несколько подробнее на поведенчески выверенных оценочных шкалах и вынужденном распределении.

- **Поведенчески выверенные оценочные шкалы (BARS)**

Поведенчески выверенные оценочные шкалы (*Behaviorally Anchored Rating Scales* - ***BARS***) представляют собой подход, разработанный П. Смитом и Л.М. Кендаллом в 1963 г. с целью создания строгой, хорошо структурированной оценочной (рейтинговой) шкалы для применения в сфере труда. Этот подход основан на методе критических случаев Дж.

Фланагана, который описывал реальные образцы (случаи) поведения на рабочем месте, релевантные успешному или неуспешному выполнению трудового задания.

На данный момент созданы две разновидности поведенчески выверенных оценочных шкал: поведенческие шкалы ожидания (*Behavioral Expectation Scales,* ***BES***) и поведенческие шкалы наблюдения (*Behavioral Observation Scales,* ***BOS***). Разработка шкал с использованием обоих подходов может требовать значительных временных затрат и привлечения множества людей, поэтому применение этих процедур обычно ограничивается крупными организациями, в которых категории работы предполагают множество должностей (или рабочих мест) в каждой.

Обе процедуры BARS начинаются одинаково. Экспертов (как правило, непосредственных руководителей или исполнителей конкретной работы) приглашают принять участие в заседаниях различных групп. В процедуре BES, первая группа экспертов описывает основные измерения, или характеристики данной работы, после чего вторая группа разрабатывает эпизоды (случаи) выполнения работы, соответствующие различным уровням каждого из выделенных первой группой измерений. Третью группу просят сделать "обратный перевод" работы первых двух групп, т. е. распределить представленные ей в случайном порядке формулировки эпизодов по соответствующим категориям.

Формулировки эпизодов, не попавшие в соответствующую категорию работы или в соответствующее измерение, отбрасываются вследствие их неопределенности. Четвертую группу просят присвоить числовые оценки оставшимся формулировкам эпизодов. И снова пункты с расхождением оценок (высокое стандартное отклонение) изымаются из набора формулировок эпизодов. Наконец, двух непосредственных руководителей просят оценить каждого своего работника, и в заключении проводится анализ на предмет того, являются ли выделенные измерения работы независимыми.

Процедура BOS также начинается с того, что группа экспертов разрабатывает описания поведенческих эпизодов. Последующие встречи группы связаны с уточнением и редактированием этих описаний, разнесением их по измерениям работы и определением того, является ли эти описания и измерения настолько независимыми, насколько это возможно. Затем кто-либо из подчиненных оценивается по каждому эпизоду или пункту с использованием шкалы, варьирующей от 7 (очень часто) к 1 (очень редко). Подобным образом можно встроить все эпизоды в шкале и определить их вклад в общую оценку.

Основным преимуществом процедур BARS, по-видимому, является вовлечение организации в процесс разработки, что обеспечивает поддержку с ее стороны при последующем внедрении и использовании шкалы.

- **Вынужденное распределение**

С помощью метода ***вынужденного распределения*** (*forced distribution*) деятельность работников оценивается по заранее заданным нормам распределения, произвольно установленным в организациях, а именно, по количеству оценочных категорий деятельности и процентным ставкам отнесения работников к каждой из них (см. рис. 8).

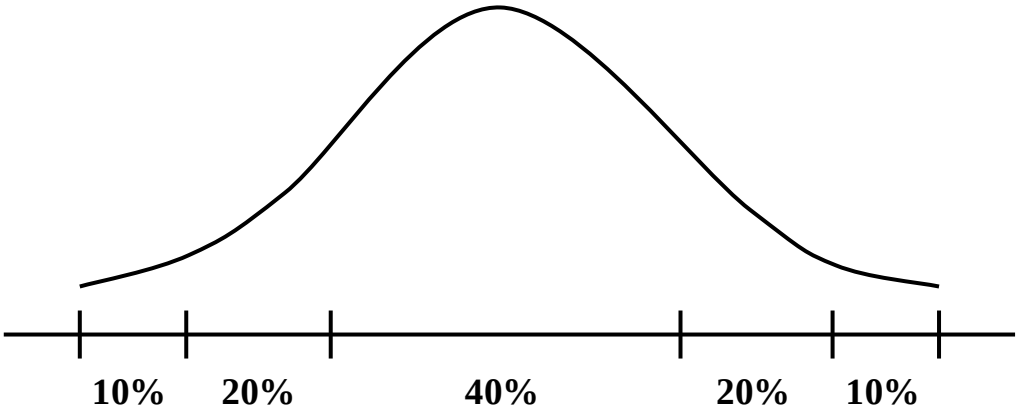


Рисунок 8. Пример графического изображения оценки деятельности работников по методу вынужденного распределения.

Оценочная система, например, может иметь пять оценочных категорий, позволяющих оценивать деятельность от “плохой” до “отличной” со следующим распределением работников (см. рис. 8):

1. оценку “плохо” получают 10% работников,
2. “ниже среднего” - 20%,
3. средний балл - 40%,
4. “выше среднего” - 20%
5. и оценку “отлично” - 10%.

Другие системы включают три оценочные категории: “плохо”, “хорошо” и “отлично”, - и процентные ставки распределения работников, соответственно, - 30%, 40% и 30%.

Применение метода вынужденного распределения аннулирует проблему выбора лучшего из имеющих равный уровень деятельности работников, встречающуюся при использовании

метода классификации. С другой стороны, распределение работников по нескольким группам ведет к однородной оценке деятельности людей, попадающих в одну оценочную категорию, не позволяет выявить отличия в их работе, что, в свою очередь, снижает ценность обратной связи и сокращает объем информации, требуемой руководителю-оценщику для принятия решений по вопросам вознаграждений. Помимо этого, нормы распределения работников по оценочным категориям могут не согласовываться с фактической ситуацией.

Для исключения ошибок центральной тенденции и ошибок снисходительности, которые уменьшают рабочую область шкалы и снижают различительную способность оценок, эффективно использование процедур ранжирования (включая парные сравнения), которые вводят принудительное различение оцениваемых лиц и, следовательно, максимизируют информацию, даваемую рейтингами (например, вынужденное распределение).

Методы статистического анализа данных, полученных от оценщиков, также направлены на повышение их надежности. Рассмотрим их.

- **Оценка компетентности экспертов (оценщиков)**

Определение компетентности оценщиков (экспертов) служит построению ранжировки объектов оценки с учетом компетентности экспертов. Компетентность экспертов оценивается по степени согласованности их оценок с групповой оценкой объектов.

Пусть нам даны экспертные оценки (см. табл. 5), которые представляют собой матрицу $A = \|a_{ij}\|_{5\times7}$: m = 1…5 – множество оцениваемых объектов, n = 1…7 – множество экспертов.

	Объект 1	**Объект 2**	**Объект 3**	**Объект 4**	**Объект 5**
Эксперт 1	1.00	3.00	3.00	4.00	5.00
Эксперт 2	2.00	2.00	3.00	5.00	4.00
Эксперт 3	4.00	2.00	5.00	2.00	3.00
Эксперт 4	3.00	3.00	1.00	4.00	5.00
Эксперт 5	2.00	1.00	2.00	3.00	5.00

Эксперт 6	2.00	3.00	4.00	4.00	5.00
Эксперт 7	3.00	3.00	2.00	1.00	5.00

Таблица 5. Экспертные оценки (исходная матрица *A*).

По исходной матрице *A* рассчитаем матрицу $C = A^T A$ (произведение транспонированной матрицы с исходной *A* – см. табл. 6).

60.00	57.00	48.00	56.00	48.00	64.00	47.00
57.00	58.00	49.00	55.00	47.00	62.00	43.00
48.00	49.00	58.00	46.00	41.00	57.00	45.00
56.00	55.00	46.00	60.00	48.00	60.00	49.00
48.00	47.00	41.00	48.00	43.00	52.00	41.00
64.00	62.00	57.00	60.00	52.00	70.00	52.00
47.00	43.00	45.00	49.00	41.00	52.00	48.00

Таблица 6. Матрица *C* (произведение транспонированной матрицы с исходной *A*).

Далее выполним следующие преобразование над строками матрицы *C*:

$$C = \begin{Vmatrix} c_{11}\, c_{12} \ldots c_{1n} \\ \tilde{n}_{21}\, \tilde{n}_{22} \ldots \tilde{n}_{2n} \\ \ldots\ldots\ldots \\ c_{n1}\, c_{n2} \ldots c_{nn} \end{Vmatrix} \Rightarrow \bar{y} = \begin{Vmatrix} \sqrt[n]{c_{11} \cdot c_{12} \cdot \ldots \cdot c_{1n}} \\ \sqrt[n]{c_{21} \cdot c_{22} \cdot \ldots \cdot c_{2n}} \\ \ldots\ldots\ldots\ldots \\ \sqrt[n]{c_{n1} \cdot c_{n2} \cdot \ldots \cdot c_{nn}} \end{Vmatrix} \Rightarrow \bar{x} = \begin{Vmatrix} \dfrac{y_1}{\sum_{i=1,n} y_i} \\ \dfrac{y_2}{\sum_{i=1,n} y_i} \\ \ldots\ldots \\ \dfrac{y_n}{\sum_{i=1,n} y_i} \end{Vmatrix}$$

В таблице 7 приведены коэффициенты компетентности экспертов (столбец *x* – взвешенные значения).

	$\bar{y}$	$\bar{x}$
Эксперт 1	53.93	0.150
Эксперт 2	52.62	0.146
Эксперт 3	48.81	0.136
Эксперт 4	53.16	0.148
Эксперт 5	45.55	0.127
Эксперт 6	59.27	0.165
Эксперт 7	46.30	0.129

Таблица 7. Расчетные коэффициенты компетентности экспертов.

- **Межэкспертная надежность - коэффициент конкордации *W***

Межэкспертная надежность (*inter-rater reliability*) - степень, в которой два или более независимых наблюдателя (эксперта) сходятся в своих оценках деятельности (поведения) индивидуумов. При измерении объектов в порядковой шкале (ранжировки, парные сравнения) для оценки межэкспертной надежности можно воспользоваться мерой согласованности мнений экспертов – ***дисперсионным коэффициентом конкордации W*** (*коэффициент согласия*), предложенным Кэндаллом:

$$W = \frac{12 \cdot S}{n^2 \cdot (m^3 - m)}$$

где m – множество оцениваемых объектов; n – множество экспертов; S рассчитывается по формуле:

$$S = \sum_{i=1}^{m} (r_m - \bar{r})^2$$

где r_m – сумма рангов для объекта m ($r_m = \sum_{j=1}^{n} r_{ij}$); $\bar{r}$ - средняя сумма рангов по всем объектам ($\bar{r} = \frac{1}{m} \cdot \sum_{i=1}^{m} r_i$).

Если ранжировки экспертов имеют связанные ранги, то коэффициент конкордации рассчитывается по следующей формуле:

$$W = \frac{12 \cdot S}{n^2 \cdot (m^3 - m) - n \cdot \sum_{j=1}^{n} T_j}$$

где T_j (показатель связанных рангов) рассчитывается по формуле:

$$T_j = \sum_{k=1}^{H_j} (h_k^3 - h_k) .$$

где h_k – объем связанных групп по каждому эксперту.

Рассмотрим на примере расчет дисперсионного коэффициента конкордации *W*.

В таблице 8 приведены оценки экспертов из таблицы 5 с учетом связанных рангов (совпавшим рангам присваивается среднее значение суммы их порядковых номеров в отсортированном ряду).

	Объект 1	**Объект 2**	**Объект 3**	**Объект 4**	**Объект 5**
Эксперт 1	1.00	2.50	2.50	4.00	5.00
Эксперт 2	1.50	1.50	3.00	5.00	4.00
Эксперт 3	4.00	1.50	5.00	1.50	3.00
Эксперт 4	2.50	2.50	1.00	4.00	5.00
Эксперт 5	2.50	1.00	2.50	4.00	5.00
Эксперт 6	1.00	2.00	3.50	3.50	5.00
Эксперт 7	3.50	3.50	2.00	1.00	5.00

Таблица 8. Оценки экспертов с учетом связанных рангов.

Расчет показателей *S*, T_j и *W* для оценок экспертов из таблицы 8 приведен в таблице 9 (H_j – число групп связанных рангов по каждому эксперту).

	$Э_1$	$Э_2$	$Э_3$	$Э_4$	$Э_5$	$Э_6$	$Э_7$	r_m	$\bar{r}$	$(r_m - \bar{r})^2$
O_1	1.0	**1.5**	4.0	**2.5**	**2.5**	1.0	**3.5**	16.0	21	25.0
O_2	**2.5**	**1.5**	**1.5**	**2.5**	1.0	2.0	**3.5**	14.5		42.25

O_3	**2.5**	3.0	5.0	1.0	**2.5**	**3.5**	2.0	19.5		2.25
O_4	4.0	5.0	**1.5**	4.0	4.0	**3.5**	1.0	23.0		4.00
O_5	5.0	4.0	3.0	5.0	5.0	5.0	5.0	32.0		121.0
H_j	1	1	1	1	1	1	1	$S = \sum_{i=1}^{m} (r_m - \bar{r})^2 = 194.5$		
h_k	$h_1 = 2$	$h_1 = 2$	$h_1 = 2$	$h_1 = 2$	$h_1 = 2$	$h_1 = 2$	$h_1 = 2$			
T_j	6	6	6	6	6	6	6	$\sum_{j=1}^{n} T_j = 42$		
$W = \frac{12 \cdot S}{n^2 \cdot (m^3 - m) - n \cdot \sum_{j=1}^{n} T_j} = \frac{12 \cdot 194.5}{7^2 \cdot (5^3 - 5) - 7 \cdot 42} = 0.418$										

Таблица 9. Пример расчета дисперсионного коэффициентом конкордации *W*.
(Жирным шрифтом выделены связанные ранги для каждого эксперта.)

Коэффициент конкордации *W* изменяется в диапазоне от 0 до 1. Значения очень близкие к нулю отражают отсутствие согласия в ранжировках суждений среди экспертов. При этом значения близкие к 1 представляют высокую степень согласия оценок между экспертами. В Приложении 6 приведены таблицы критических значений для проверки значимости коэффициента конкордации *W*.

- **Согласованность оценок - коэффициенты корреляций результатов оценщиков**

Для проверки согласованности оценок могут использоваться методы ранговой корреляции, которые позволяют определить зависимости между ранжировками пар экспертов. Наиболее часто для этих целей используется ранговый коэффициент корреляции Спирмена (см. Приложение 2). Рассчитаем коэффициент корреляции Спирмена для данных "Эксперта 1" и "Эксперта 2" из таблицы 5 (расчеты представлены в таблице 10). Поскольку ранжировки экспертов имеют связанные ранги, то коэффициент корреляции Спирмена вычисляется по формуле:

$$r_s^{'} = \frac{r_s - T_x - T_y}{\sqrt{(1 - 2 \cdot T_x) \cdot (1 - 2 \cdot T_y)}},$$

где $T_x = \frac{1}{2 \cdot (m^3 - m)} \cdot \sum_{k=1}^{H_x} (h_k^3 - h_k)$, $T_y = \frac{1}{2 \cdot (m^3 - m)} \cdot \sum_{k=1}^{Hy} (h_k^3 - h_k)$.

Объекты	**Эксперт 1**	**Эксперт 2**	d_i	d^2_i

1	1.0	**1.5**	–0.5	0.25
2	**2.5**	**1.5**	1.0	1.00
3	**2.5**	3.0	–0.5	0.25
4	4.0	5.0	–1.0	1.0
5	5.0	4.0	1.0	1.0
H_j	1	1	$\sum d^2_i = 3.5$	
h_k	$h_1 = 2$	$h_1 = 2$	$r_s = 1 - \frac{6 \cdot 3.5}{5 \cdot (25-1)} = 0.825$	
$\sum_{k=1}^{H_j} (h_k^3 - h_k) =$	6	6		
$T_i =$	$\frac{6}{2 \cdot (5^3 - 5)} = 0.025$	$\frac{6}{2 \cdot (5^3 - 5)} = 0.025$	$r_s' = \frac{0.825 - 0.025 - 0.025}{\sqrt{(1 - 2 \cdot 0.025) \cdot (1 - 2 \cdot 0.025)}} = 0.816$	

Таблица 10. Пример расчета рангового коэффициент корреляции Спирмена.
(Жирным шрифтом выделены связанные ранги для каждого эксперта.)

ТРЕБОВАНИЯ К ВЫБОРКЕ ИСПЫТУЕМЫХ ПРИ ИЗУЧЕНИИ НАДЕЖНОСТИ

- **Количественные требования**

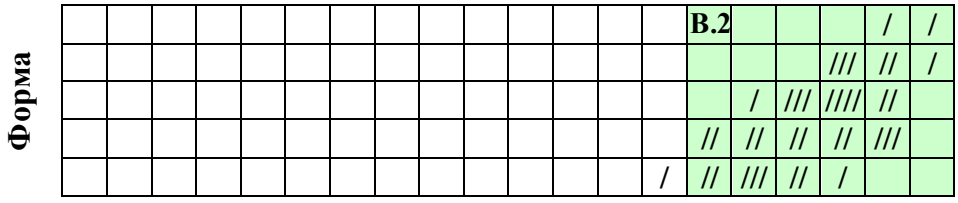
Поскольку, как и любая другая статистическая величина, стандартная погрешность коэффициента корреляции, используемого при оценке надежности, связана с объемом выборки, на которой она была получена, то вполне естественно, что должны использоваться большие выборки, чтобы минимизировать погрешность такого рода. П. Клайн исследовал стандартные погрешности корреляций и пришел к выводу, что с выборкой из 200 испытуемых этот источник погрешностей уже можно не принимать в расчет. Таким образом, он рекомендует для исследования надежности тестов выборки с объемом не менее 200, хотя и желательны большие объемы. Для точности вычислений по формуле К-R20, в которой используется процент от количества испытуемых, давших ключевые ответы, необходимы, с его точки зрения, большие выборки, и 200, в данном случае, - это лишь желательный минимум.

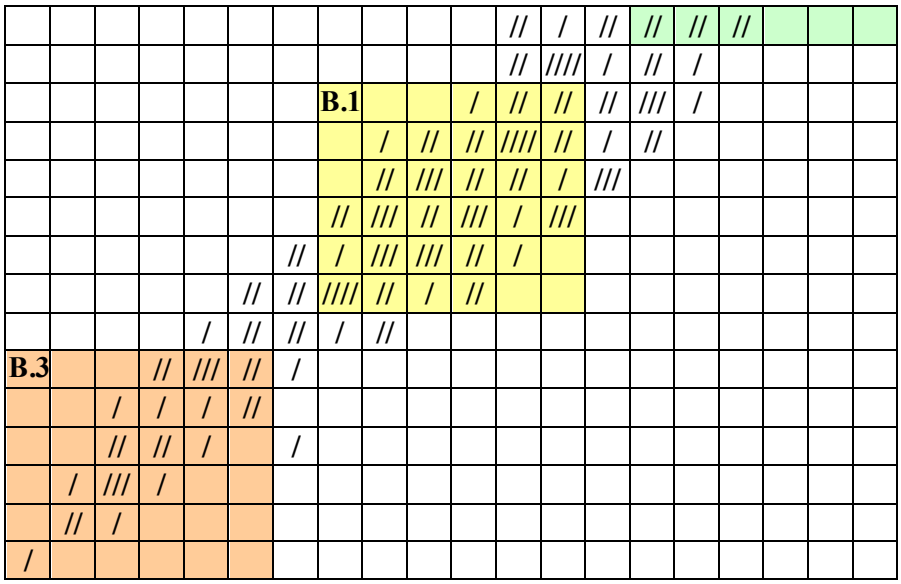
- **Содержательные требования**

На характеристику надежности методик большое влияние оказывает исследуемая выборка. Она может как снижать, так и завышать этот показатель. Например, надежность может быть искусственно занижена, если в выборке небольшой разброс результатов, т.е. если результаты по своим значениям близки друг к другу. Так выборка В.1 на рисунке 9 попадает в небольшой прямоугольник в центре графика. Очевидно, что коэффициент корреляции между результатами выполнения двух форм теста для данной выборки будет близок к нулю. Индивиды, попадающие в данную выборку представляют собой весьма гомогенную группу по результатам выполнения теста. Неоправданное завышение надежности может возникнуть при анализе результатов выборки, состоящей из группы, имеющей очень высокие результаты (см. рис. 9, В.2), и из группы с очень низкими оценками по тесту (см. рис. 9, В.3). Тогда эти далеко отстоящие друг от друга результаты не будут перекрываться, даже если и вмешаются в условия эксперимента случайные факторы.

В настоящее время надежность все чаще определяется на наиболее однородных выборках, т.е. на выборках, схожих по полу, возрасту, уровню образования, профессиональной подготовке и т.п. Для каждой такой выборки приводятся свои коэффициенты надежности. Приводимый показатель надежности применим только к группам, подобным тем, на которых он определялся. Если методика применяется к выборке, отличающейся от той, на которой проверялась ее надежность, то эта процедура должна быть проведена заново.

Согласно П. Клайну, для теста, предназначенного к использованию в разных группах, часто полезно показать, что он надежен для каждой группы в отдельности. Здесь допустимы объемы выборок ниже 200. Например, если установлена устойчивая, высокая надежность для выборок из 100 "белых воротничков", 100 "синих воротничков" и 100 "руководителей", тогда мы можем быть уверены, что он будет надежен в этих группах. Автор обращает внимание, что одна конкретная выборка из 300 испытуемых не была бы репрезентативной для каждой из этих групп.





Форма теста А

Рисунок 9. Влияние выборки на оценку надежности теста.

ОБЩИЙ ОБЗОР ТИПОВ И КОЭФФИЦИЕНТОВ НАДЕЖНОСТИ

Различные виды рассмотренных коэффициентов надежности сведены в таблицы 11 и 12. В таблице 11 методы, применяемые для оценки каждого типа надежности, сгруппированы в зависимости от числа требуемых для этой цели форм теста и сеансов тестирования. В таблице 12 представлены источники дисперсии, трактуемые каждым из методов как дисперсия ошибок.

Необходимое число сеансов тестирования	**Необходимое число форм теста**	
	одна	**две**
один	• Метод расщепления на эквивалентные половины (формулы Спирмена-Брауна, Фланагана, Кристофа, Рюлона) • Метод Кьюдера-Ричардсона (дихотомическая шкала ответов) и коэффициент	• Метод взаимозаменяемых форм – непосредственный (коэффициенты корреляции)

Необходимое число сеансов тестирования	Необходимое число форм теста	
	одна	**две**
	альфа Кронбаха (универсальная шкала ответов)	
два	• Метод «тест - ретест» (коэффициенты корреляции)	• Метод взаимозаменяемых форм отсроченный - (коэффициенты корреляции)

Таблица 11. Классификация методов измерения надежности в зависимости от требуемого числа форм теста и сеансов тестирования.

Вид коэффициента надежности	Дисперсия ошибок
• Ретестовый	• Временная выборка
• Взаимозаменяемых форм (непосредственный)	• Выборка содержания
• Взаимозаменяемых форм (с временным интервалом)	• Временная выборка и выборка содержания
• Эквивалентных половин теста	• Выборка содержания
• Кьюдера-Ричардсона и альфа Кронбаха	• Выборка содержания и неоднородность содержания
• Оценщика	• Различия между оценщиками

Таблица 12. Источники дисперсии ошибок, связываемые с коэффициентами надежности.

ОБЩИЕ ПРИНЦИПЫ ДЛЯ ИНТЕРПРЕТАЦИИ КОЭФФИЦИЕНТОВ НАДЕЖНОСТИ

Как было сказано выше, о надежности теста мы судим по коэффициенту надежности (r_t), который выражается числом в диапазоне 0÷1.00: r_t = 0 указывает об отсутствии надежности, и r_t = 1.00 – о совершенной надежности. В разных источниках можно

обнаружить различные рекомендации по интерпретации значений коэффициентов надежности. Например, в [21] приведена следующая таблица для интерпретации коэффициентов надежности (см. табл. 13):

Величина коэффициента надежности	**Интерпретация**
0.90 и выше	отличная
0.80 ÷ 0.89	хорошая
0.70 ÷ 0.79	достаточная
ниже 0.70	Тест может иметь ограничения в применимости

Таблица 13. Общие принципы для интерпретации коэффициентов надежности.

Таблицу 13 можно также рекомендовать при интерпретации коэффициента надежности, рассчитанного по формуле *KR-20*.

В учебном пособии под редакцией К.М. Гуревича и Е.М. Борисовой [15] методика признается надежной, когда полученный коэффициент имеет значение не ниже 0.75÷0.85. Лучшие по надежности тесты дают коэффициенты порядка 0.90 и более.

Л.Ф. Бурлачук и С.М. Морозов [6] утверждают, что в большинстве применяемых методик редко удается получить значения коэффициентов надежности, превышающие 0.7÷0.8. Опираясь на практику психологической диагностики они полагают, что при использовании формул Кьюдера-Ричардсона и альфа (оценка внутренней согласованности) тест можно считать надежным, если $r_t \geq 0.6$.

Согласно А. Анастази и С. Урбина [3] коэффициенты надежности обычно превышают значение 0.80 и даже 0.90. Из этого делается вывод, что желательным может служить коэффициент надежности, превышающий значение 0.8.

В своем "Справочном руководстве" [7] П. Клайн утверждает, что наименьшим удовлетворительным значением для ретестовой надежности является значение $r_t = 0.7$. В тоже время, как отмечает редактор этой книги Л.Ф. Бурлачук, указанный предельный

коэффициент надежности в известной мере можно считать условным, и для некоторых тестов личности показатель ретестовой надежности может быть ниже, при этом диагностическая ценность методики не снижается.

Кондаков И.М. с соавторами отмечают [10], что при оценке гомогенности заданий или внутренней устойчивости теста коэффициентом альфа Кронбаха, профессионально разработанные важные тесты должны иметь внутреннюю устойчивость на уровне не менее 0.90. Менее важные стандартизированные тесты (которые проводятся однажды, например, при приеме на работу) должны иметь значение коэффициента не менее 0.80. Если окончательная оценка субъекта основывается на нескольких измерениях (тестах, деловых играх, собеседовании), то для теста желательно иметь значение надежности выше 0.70.

Эти авторы предлагают следующую общую трактовку значений коэффициента надежности (см. табл. 14):

Величина коэффициента надежности	**Интерпретация**
0.90 и выше	**Высокая надежность** *Необходима в случаях, когда:* • на данных теста предполагается делать серьезные выводы; • экзаменуемые разделены на множество разных категорий на основании относительно небольших индивидуальных различий, например, интеллекта.
0.80 и выше	**Средняя или высокая надежность** (примерно $100 \times 0.8^2 = 16\%$ изменчивости в тестовом балле приходится на долю ошибки).
около 0.70	**Низкая надежность** *Приемлема, если:* • тест используется для получения предварительных выводов; • тест используется для сортировки людей на небольшое количество групп на основании больших индивидуальных различий, например, роста или интровертности-экстравертности.
менее 0.60	**Неприемлемо низкая надежность**

Таблица 14. Общая трактовка коэффициентов надежности.

Необходимо помнить, что ретестовая надежность может быть невысокой в силу динамичности измеряемого конструкта. Примером может служить вариабельность сердечного ритма как показатель функционального состояния человека. При этом валидность теста остается высокой.

Коэффициент надежности обладает доверительным интервалом, определение которого особенно важно в связи с большим количеством факторов, способных влиять на его значение [6]. Доверительный интервал (E_{rt}) для r_t определяется по формуле:

$$E_{rt} = Z_{(r)} \pm Z_{\hat{e}\eth\grave{e}\grave{o}} \cdot \sigma_{rt}$$

где σ_{rt} - стандартная ошибка коэффициента надежности ($\sigma_{rt} = \frac{1}{n-3}$); $Z_{(r)}$ – Z-преобразование Фишера = $\frac{1}{2}\ln\frac{1+r_t}{1-r_t}$ (определяется по статистическим таблицам).

На практике принимается во внимание только нижняя граница r_t ($Z_{крит}$ при γ = 0.05 составляет 1.96, при α = 0.01 $Z_{крит}$ = 2,58).

Надежность теста можно выразить через ***стандартную ошибку измерения*** (***SEM*** - сокр. от *Standard Error of Measurement*), называемую также стандартной погрешностью измерения или стандартной ошибкой показателя. Она задает границу ошибки, которую можно ожидать в индивидуальных оценках по тесту, поскольку надежность теста не совершенна. *SEM* представляет степень достоверности (уверенности), что "истинная" оценка индивида лежат внутри определенного диапазона оценок. Эта мера особенно удобна для интерпретации индивидуальных показателей. Поэтому для многих целей тестирования она более полезна, чем коэффициент надежности. Зная коэффициент надежности теста, стандартную ошибку измерения легко вычислить по следующей формуле:

$$SEM = SD_t \cdot \sqrt{1-r_t},$$

где SD_t - стандартное отклонение показателей теста; r_t - коэффициент надежности (оба вычисленные на одной группе).

Например, если стандартные показатели по конкретному тесту на внимание имеют SD_t =15 и коэффициент надежности r_t =0.89, то для данного теста стандартная погрешность измерения будет равна:

$$SEM = 15 \cdot \sqrt{1-0.89} \approx 5.$$

Допустим, было выполнено 100 измерений уровня внимания у отдельного индивида. Вследствие разного рода случайных ошибок, которые влияют на надежность теста, результаты тестирования варьируются вокруг истинного показателя индивида, подчиняясь нормальному распределению. Среднее этого распределения ста показателей можно принять за «истинный показатель» для данного использования теста, а стандартное отклонение - за соответствующую стандартную ошибку измерения (*SEM*). Как и любое стандартное отклонение, стандартную ошибку можно интерпретировать в единицах плотности нормального распределения. При нормальном распределении в интервал $M \pm 1\ \sigma$ попадает приблизительно 68% всех случаев. Следовательно, имеется примерно 2 шанса против 1 (точнее, 68:32), что показатели индивида по тесту на внимание будут колебаться в пределах ± 1 SEM или 5 единиц (для нашего примера) в обе стороны от ее истинного значения уровня внимания. Если истинный уровень внимания = 110, можно ожидать, что в 2/3 (68%) случаев показанные индивидом результаты попадут в интервал между 105 и 115.

Когда мы хотим чувствовать себя увереннее в наших предсказаниях, мы можем выбрать более высокое соотношение шансов, чем 2:1. Так интервал $M \pm 3\sigma$ покрывает 99.7% случаев. Обратившись к таблицам плотности нормального распределения, можно удостовериться, что интервал $M \pm 2.58\sigma$ включает точно 99% случаев. Следовательно, имеется 99 шансов против 1, что уровень внимания индивида попадет в интервал с границами, отстоящими на 2.58 *SEM* или на $2.58 \times 5 = 13$ единиц в обе стороны от истинного уровня внимания. Таким образом, можно с 99% степенью уверенности (1 шанс ошибиться против 100) утверждать, что уровень внимания индивида при любом одиночном проведении этого теста будет лежать в пределах значений от 97 до 123 (110 ± 13)[3]. Если бы индивиду предъявили 100 эквивалентных тестов, то его уровень внимания мог бы выйти за границы этой области значений только однажды.

На практике мы не располагаем истинными показателями; обычно в нашем распоряжении имеются лишь показатели, полученные при одном-единственном проведении теста. В этих обстоятельствах мы можем применить выше приведенные рассуждения в обратном порядке. Если маловероятно, что полученный тестируемым показатель отклонится от его истинного значения более чем на 2.58 *SEM*, мы могли бы утверждать, что его истинное

[3] П. Клайн подчеркивает, что стандартная погрешность измерения может быть использовано для определения доверительных границ полученных показателей, но надо помнить, что эти зоны располагаются симметрично вокруг истинного, а не эмпирического (результат конкретного замера) показателя.

значение должно лежать в пределах 2.58 *SEM* от полученного им результата. Хотя нельзя установить вероятность справедливости этого утверждения для любого отдельного показателя, полученного конкретным испытуемым, можно сказать, что оно будет верным для 99% всех возможных случаев. Следуя этому рассуждению, Галликсен (Gulliksen) предложил использовать стандартную ошибку измерения для оценки разумных границ истинного показателя у лиц с любым полученным в единичном измерении показателем. *SEM* являет полезным измерением точности индивидуальных тестовых оценок. Чем меньше *SEM*, тем более точные измерения.

Используя стандартную ошибку измерения, Л.Ф. Бурлачук и С.М. Морозов [6] предлагают следующую формулу для расчета истинного значения тестового балла субъекта:

$$x_t = r_t \cdot x_i + \bar{x} \cdot SEM^2 = r_t \cdot x_i + \bar{x} \cdot (1 - r_t)$$

где x_t - истинное значение тестового балла; x_i - эмпирический балл испытуемого (результат по тесту); r_t - коэффициент надежности теста; $\bar{x}$ - среднее значение оценок по тесту.

Например, у испытуемого при обследовании по шкале Векслера оценка вербального интеллектуального показателя определена в 107 баллов. Среднее значение $\bar{x}$ для шкалы составляет 100, а надежность тестовой шкалы r_t = 0.89. При этом истинное значение будет равно:

$$x_t = 0.89 \times 107 + 0.11 \times 100 = 106.2.$$

Как мы помним, любой коэффициент надежности можно интерпретировать непосредственно в процентах дисперсии показателей, приписываемой разным источникам (см. Общие понятия надежности теста). Для этого несколько видоизменим формулу относительной доли дисперсии ошибки (S_o^2):

$$S_o^2 = (1 - r_t) \cdot 100\%$$

где r_t - надежность теста.

Тогда коэффициент надежности 0.85 означает, что 85% дисперсии показателей теста зависят от истинной изменчивости (дисперсии) измеряемой черты, а 15% - от дисперсии ошибок.

ВАЛИДНОСТЬ ТЕСТА

ОБЩИЕ ПОНЯТИЯ

- **Классификация основных типов валидности теста**

Валидность теста - это комплексная характеристика, включающая (см. рис. 10), с одной стороны, сведения о том, пригодна ли методика для измерения того, для чего она была создана (*предмет измерения*), а с другой стороны, какова ее эффективность и практическая ценность (*цель измерения*).

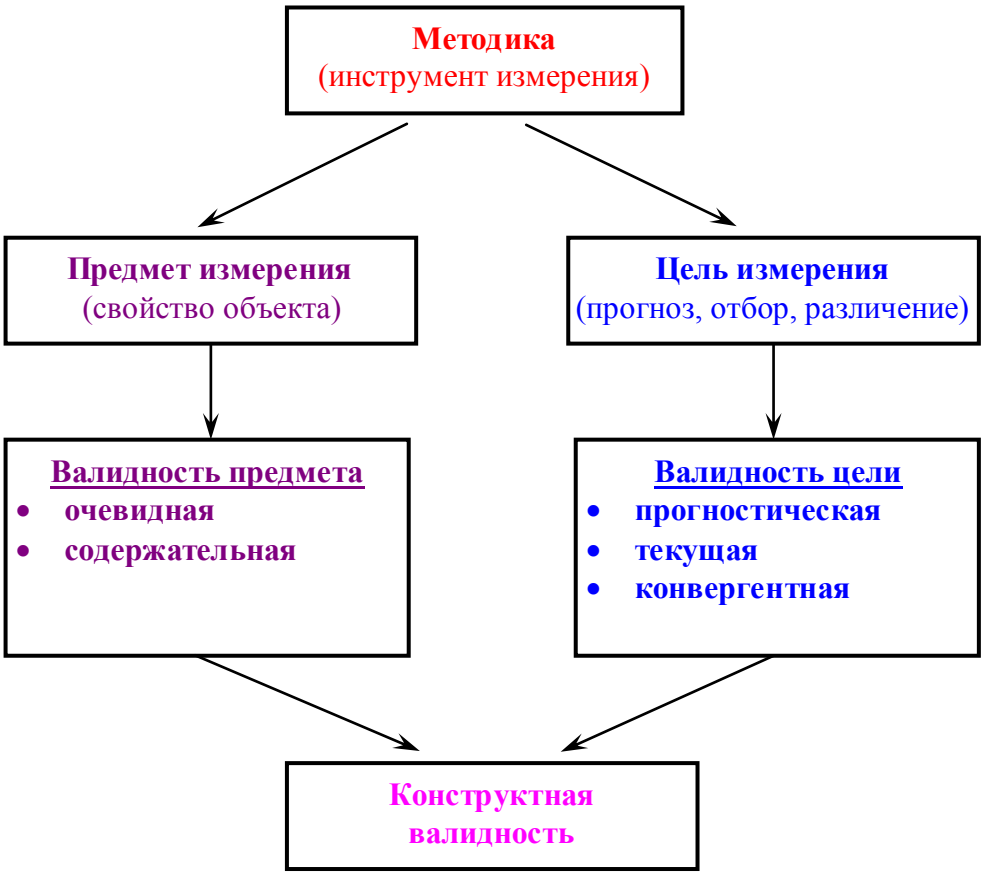


Рисунок 10. Основные типы валидности теста.

При оценке валидности предмета измерения нас интересует то свойство (характеристика объекта), для измерения которого методика была разработана, насколько хорошо она это делает (***теоретическая валидизация теста***). В этом контексте тест называется валидным, если он измеряет именно то свойство, для измерения которого он предназначен.

При валидизации цели измерения главный акцент делается на связи результатов тестирования с определенными областями практики: какие выводы можно сделать из полученных по тесту показателей (***прагматическая валидизация теста***). В этом контексте тест называется валидным, если его результаты позволяют эффективно решать вопросы прогноза, отбора, дифференциации.

- **Основные факторы, влияющие на валидность теста**

Прежде чем перейти к оценкам различных типов валидности, остановимся дополнительно на факторах, которые могут влиять на валидность теста.

Во-первых, это *мотивационные искажения*. При использовании тестов можно выделить ситуацию "экспертизы" и ситуацию "клиента. В ситуации "экспертизы" тест проводится с целью определения возможностей, способностей обследуемого к выполнению, например, определенной деятельности. Результатами, по существу, распоряжается не человек, который проходит тест, а заказчик тестирования (например, организация, осуществляющая отбор персонала). В ситуации "экспертизы" достоверность полученных результатов может снижаться из-за излишнего перенапряжения или перевозбуждения обследуемых (диагностика психических процессов), или повышенной "закрытости" (личностная диагностика), вызванные мотивационными влияниями. В ситуации "клиента" человеком, проходящим тест, движет мотив самопознания, он сам распоряжается своими результатами, а потому мотивационные искажения будут не столь сильными.

Социальная желательность - важный фактор искажений ответов на личностные опросники. Человек живет в обществе и ограничен в своих возможностях демонстрировать неодобряемые обществом модели поведения. Иногда это выражается в предпочтении ответов теста, соответствующих более одобряемому поведению.

Еще одним фактором снижения достоверности результатов в случае психологических тестов может быть *осведомленность испытуемого об измеряемом психическом свойстве*. Иными словами, зная «на что именно его ловят», человек может «перехитрить» тест. Поэтому в инструкциях испытуемым к тестам направленность теста обычно обозначают общими словами «Тест направлен на выявление некоторых особенностей Вашего характера».

В заключении необходимо заметить, что низкая надежность теста также является источником его плохой валидности, и рассмотренные ранее факторы влияющие на надежность теста (см. Общие понятия надежности теста) справедливы, таким образом, и для валидности. Чтобы быть валидным, тест должен быть надежен, но надежность не гарантирует валидность. Высокая надежность означает, что тест измеряет какое-то свойство очень точно, но какое именно – остается под вопросом. Решение его и является задачей валидизации теста.

ВАЛИДИЗАЦИЯ ПО ПРЕДМЕТУ ИЗМЕРЕНИЯ (ВАЛИДНОСТЬ ПО СОДЕРЖАНИЮ)

- **Очевидная (внешняя) валидность**

Очевидная валидность *(face validity)* - это возникающие у испытуемых представления о тесте (о его содержании, целях тестирования, сфере применения, прогностической ценности и т.п.). Очевидная (или внешняя) валидность выступает в роли фактора, побуждающего испытуемых к сотрудничеству в ситуации обследования, способствует более серьезному и ответственному отношению к работе при выполнении заданий и к восприятию заключений, формулируемых психологом. Если содержание вопросов или заданий теста крайне далеки от профессиональной реальности испытуемых, воспринимаются ими как примитивные или оскорбительные, то это может вызвать лишь нежелание испытуемых выполнять поставленные перед ними задачи и стремление всячески саботировать цели обследования.

Для определения очевидной валидности достаточно просто опросить испытуемых (экспертов), принимающих участие в процедуре оценки и отбора заданий для теста, представляются ли они им хорошим средством измерения данной переменной или нет, являются ли они адекватными поставленной цели измерения.

Это довольно смутная процедура валидизации теста, которая обычно используется только на начальных стадиях построения теста.

- **Содержательная валидность**

Содержательная валидность *(content validity)* характеризует степень соответствия содержания заданий теста той реальной деятельности, в которой проявляется измеряемое психическое свойство, и основывается на детальном исследовании содержания пунктов теста. Она имеет большое значение для тестов, исследующих деятельность, близкую или совпадающую с реальной. Например, это важно для тестов учебных и профессиональных достижений, тестов математических, музыкальных способностей, словарного запаса и знаний грамматики. Содержательная валидность представляет интерес в основном при конструировании тестов, когда должен быть точно определен материал, используемый для тестирования. Часто изучаемая деятельность носит, как правило, синтетической характер, складывается из многих, подчас разнородных, факторов (проявления личностных особенностей, комплекс необходимых знаний, умений и навыков, специальные способности и т.д.). Поэтому одной из важнейших задач создания адекватной модели тестируемой деятельности является подбор таких заданий, которые будут охватывать главные аспекты изучаемого феномена в правильной пропорции к реальной деятельности в целом. Если можно показать, что задания теста отражают все аспекты исследуемой области поведения, то тест является, по существу, содержательно валидным.

Например, чтобы тест математических способностей имел достаточный уровень содержательной валидности, его пункты не должны иметь таких формулировок, при которых для человека, выполняющего тест, решающими оказываются вербальные способности, необходимые для того, чтобы понять, о чем спрашивается в этом пункте. Далее, содержание должно быть уравновешено таким образом, чтобы все тестируемые аспекты были представлены соответственно; тест не должен быть перегружен, скажем, пунктами на умножение в ущерб пунктам на сложение.

Содержательная валидность закладывается в тест уже при подборе заданий и основывается на мнениях "экспертов" относительно уместности используемых материалов:

1. Первым этапом валидизации является определение круга исследуемых свойств и видов деятельности, расчленение сложной способности или деятельности на элементы.
2. На втором этапе разрабатывают собственно модель тестовой деятельности на основе наиболее важных элементов реальной деятельности. Для этого могут быть использованы эксперты в данной области, которые указывают, какой материал они считают существенно важным.

3. Наконец, на последнем этапе материал преобразуется в задания теста и проводится анализ степени соответствия разработанной модели реальной деятельности, проверка соответствия пропорций представленности элементов в заданиях теста и в реальной деятельности. Эти задания опять направляются экспертам, чтобы посмотреть, не обнаружат ли они каких-либо грубых упущений или заданий, дублирующих друг друга.

Приведем пример обеспечения содержательной валидности при конструировании Кэттеллом личностного теста 16 PF:

(1) Просмотр словаря в поисках всех терминов, описывающих поведение;

(2) Избавление от тех терминов, которые эксперты сочли синонимами;

(3) Ранжирование испытуемых по оставшимся описаниям и выделение кластеров;

(4) Формулирование заданий, предназначенных для выявления этих кластеров.

Это был тщательно разработанный метод исследования всей генеральной совокупности переменных и попытки обеспечить содержательную валидность, который потребовал огромных денежных и временных ресурсов и не рекомендуется обычному разработчику тестов. Согласно П. Клайну, при конструировании тестов личности и мотивов, если нет ясных описаний предмета измерения, рассмотрение содержательной валидности неуместно.

П. Клайн приводит следующую общую процедуру для определения содержательной валидности.

Тесты достижений:

(1) Укажите точно категорию лиц, для которых этот тест предназначен.

(2) Определите навыки, подлежащие тестированию, возможно, после их анализа.

(3) Передайте этот список экспертам в данной области (учителям и т.п.) для проверки, нет ли упущений.

(4) Преобразуйте этот список в перечень заданий, используя, когда это возможно, равное количество заданий на каждый навык.

(5) Представьте эти задания экспертам для проверки.

(6) Подвергните задания обычным процедурам конструирования тестов. В результате должен быть получен содержательно валидный тест.

Другие тесты:

(1) Если существует литература с описаниями, просмотрите ее и преобразуйте описания в особенности поведения.

(2) Для каждой упомянутой особенности поведения сформулируйте ряд заданий.

(3) Когда литература с описаниями отсутствует, получите описания поведения от грамотных специалистов; например, для изучения зависимости инфантильных пациентов опросите их лечащих врачей и медицинских сестер с целью получить описание зависимого поведения их пациентов.

(4) Как и на шаге (2) выше, преобразуйте описания в задания.

(5) Подвергните задания теста обычным процедурам конструирования теста.

Отмечается растущий интерес к применению содержательной валидизации тестов для отбора персонала. Во всех своих формах такая валидизация опирается на систематический ***анализ содержания работы*** *(job analysis)*, который рассмотрен в Приложении 5.

ВАЛИДИЗАЦИЯ ПО ЦЕЛИ ИЗМЕРЕНИЯ (ВАЛИДНОСТЬ ПО КРИТЕРИЮ)

Для проверки валидности цели измерения выбирается какой-нибудь независимый от методики внешний критерий, определяющий успех в той или иной деятельности (учебной, профессиональной и т.п.), и с ним сравниваются результаты диагностической методики. Если связь между ними признается удовлетворительной, то делается вывод о практической эффективности, действенности диагностической методики. Это имеет большое значение при решении вопросов отбора. Данный класс валидности называют ***критериальным***, он включает в себя такие типы валидности как ***прогностическая, текущая, конвергентная***, и представляет собой комплекс характеристик, отражающий соответствие результатов тестирования определенным значениям критериальной переменной или вероятности критериального события. В качестве критерия выступают либо независимые от результатов теста непосредственные меры исследуемого качества (такие как уровень достижения в какой-либо деятельности, степень развития способности, выраженность определенного свойства личности и т.д.), либо показатели социально или производственно-значимых результатов деятельности

(производительность труда в индустриальной психологии, успеваемость в педагогической психологии, устойчивость брака в психологии семьи и т.п.). Валидность по критерию наиболее уместна при изучении локальной валидизации, при которой оценивается эффективность теста для конкретной программы тестирования, например, когда какая-либо фирма намерена оценить тест для отбора поступающих к ним на работу или когда какой-либо учебный центр хочет установить, насколько пригоден тест способности к обучению для предсказания успешного освоения обучаемыми материала данного курса. Валидность по критерию лучше всего называть практической валидностью теста при локальном применении.

- **Основные типы критериев**

Американские исследователи Тиффин и Маккормик, проведя анализ внешних критериев, используемых для доказательства валидности цели измерения, выделили четыре их типа:

1) **критерии исполнения** (в их число могут входить такие, как результаты выполнение реальной деятельности, результаты специального обучения, время, затраченное на подготовку к новой должности, темп должностного роста и квалификации и т.п.);

2) **субъективные критерии** ("экспертные оценки": они включают различные виды ответов, которые отражают отношение, например, инструктора или руководителя к обучаемому или подчиненному, его мнение, взгляды, предпочтения; обычно субъективные критерии получают с помощью интервью, опросников, анкет);

3) **физиологические критерии** (они используются при изучении влияния окружающей среды и других ситуационных переменных на организм и психику человека; замеряется частота пульса, давление крови, электросопротивление кожи и т.д.);

4) **критерии случайностей** (применяются, когда цель исследования касается, например, проблемы отбора для работы таких лиц, которые менее подвержены несчастным случаям).

В научных исследованиях часто применяются специальные лабораторные критерии. Например, для создания компактного тест-опросника на тревожность в качестве критерия его валидности может быть разработан специальный трудоемкий объективный лабораторный эксперимент, в котором воспроизводится реальная ситуация тревожности.

В критериях исполнения различают *промежуточные* и *конечные критерии*. При разработке теста для отбора курсантов военных летных училищ или теста медицинских способностей, например, конечными критериями были бы выполнение боевых заданий летчиком и достижение положительных результатов практикующим врачом соответственно. Очевидно, для получения таких критериальных данных потребовалось бы много времени. Сомнительно к тому же, что в реальной деятельности вообще можно получить действительно конечный критерий. Даже если бы такой конечный критерий в итоге оказался в нашем распоряжении, он, вероятно, подвергался действию множества неконтролируемых факторов, что снижало бы ценность результатов. По этим причинам в качестве критериальных мер часто используются такие промежуточные критерии, как данные о результативности обучения на той или иной стадии.

Наилучшие во многих отношениях меры критерия валидизации основаны на последующем ***выполнении реальной деятельности*** (*job performance*). В какой-то мере эти критерии использовались при валидизации тестов общего интеллекта и личности, но в значительно большей степени — при валидизации тестов специальных способностей. Кроме того, они обычно применяются для валидизации изготавливаемых по особому заказу тестов, касающихся отбора кадров для профессий, входящих в специальный перечень (авиадиспетчеры, операторы АЭС и т. д.). Большинство показателей выполнения профессиональной деятельности, не являясь, вероятно, конечными критериями, обеспечивают по крайней мере надежные промежуточные критерии для многих целей тестирования. В этом отношении они предпочтительнее данных о прохождении специального обучения. Вместе с тем при измерении выполнения той или иной работы не удается в такой степени стандартизовать условия, как в случае профессионального обучения. Более того, поскольку в этом случае требуется более длительный контроль за работающими, использование критерия выполнения реальной деятельности, вероятно, влечет за собой сокращение выборки валидизации. Ввиду того, что работники, занимающие номинально одинаковые должности, в разных организациях выполняют фактически неодинаковые функции, в руководстве к тесту вместе с данными о валидности относительно критерия реальной деятельности следует указать не только использованные при валидизации конкретные показатели этого критерия, но и дать краткую характеристику обязанностей, выполнявшихся этими работниками.

Для валидизации тестов профессионального отбора большое значение играют субъективные оценки или ***рейтинги*** (*ratings*). Примером могут служить оценки

эффективности деятельности персонала АЭС по моделям компетенций для валидизации тестовых процедур методом контрастных групп (текущая или прогностическая валидность). Высокая полезность субъективных оценок объясняется огромными трудностями в установлении объективных критериев в данной области. Хотя эти оценки не свободны от ошибок, свойственных всем субъективным суждениям, они представляют собой ценный источник критериальных данных при условии их получения в тщательно контролируемых условиях.

Одним из условий, влияющих на рейтинговые оценки, является *степень релевантной связи (relevant contact)* оценщика с оцениваемым человеком. Недостаточно просто долгое время знать человека; оценщик должен иметь возможность наблюдать его в таких ситуациях, где могло проявиться изучаемое поведение. Например, если у работника не было возможности принимать решения в ходе выполнения своих функциональных обязанностей, эта способность не может оцениваться его непосредственным руководителем. Во многих ситуациях проведения рейтингов желательно оставить время для проверки оценок другими способами, если оценщик не имел возможности наблюдать конкретное свойство у данного человека.

Другие способы повышения точности субъективных оценок и сокращения общих типов ошибок были нами рассмотрены ранее (см. Ошибки оценщика).

Другой оценочной процедурой, особенно полезной при сборе мнений равных по положению людей, является ***методика выдвижения кандидатур*** *(nominating technique).* Впервые разработанная в социометрии (J.L. Moreno) для исследования структуры группы, эта методика может использоваться в любой группе лиц, которые находились вместе достаточно долго, чтобы познакомиться друг с другом. Каждого человека просят выбрать одного или более членов группы, с которым он хотел бы взаимодействовать в сложной производственной ситуации, совместно пройти специальную переподготовку, общаться после работы или выполнять любую другую из перечисляемых в методике функций. Респондентов можно попросить выбрать столько членов группы, сколько они пожелают, или же назвать их в определенном порядке (первый, второй, третий выбор), или указать только одно лицо для каждой функции.

Когда эта методика используется для индивидуальной оценки, количество выборов, полученных любым отдельным человеком, может помочь распознать потенциальных

лидеров (получивших много выборов), равно как и членов группы, оказавшихся в изоляции (редко или совсем не упоминавшихся). В дополнение можно рассчитать ряд индексов для более точной оценки каждого члена группы. Проще всего подсчитать сколько раз человека выбрали для выполнения конкретной функции, что можно трактовать как его внутригрупповую оценку. Методику выдвижения кандидатов можно применять по отношению к любому интересующему нас аспекту поведения. Например, респондентов можно попросить назвать человека с наиболее оригинальными идеями, или человека, на которого можно положиться в работе, или лучшего спортсмена. Кроме того, респондентов можно попросить назвать не только того, кто больше всего соответствует данной характеристике, но и того, кто менее всего соответствует ей. В последнем случае при подсчете суммарного показателя каждого члена группы положительным выборам можно было бы приписать весовой коэффициент +1, а отрицательным - весовой коэффициент –1. В тех случаях, когда допускаются отрицательные номинации, следует быть особенно осторожным, чтобы предотвратить любые потенциально вредные воздействия методики выдвижения кандидатур на участников исследования. Следует добавить, что оценки индивидуума членами его круга можно получить и другими способами, такими как ранжирование или парные сравнения, но методика выдвижения кандидатур, по-видимому, оказалась наиболее успешной и потому используется чаще других.

Независимо от способа, социометрическое оценивание индивидуума членами его круга обычно выделяется как одна из самых надежных методик получения оценок в столь различных группах, как военнослужащие, руководители среднего звена на промышленных предприятиях, волонтеры Корпуса мира, школьники и студенты. При проверке относительно разнообразных практических критериев межличностных отношений, такие оценки обнаружили хорошую текущую и прогностическую валидность.

В заключение заметим, что в качестве внешних критериев при оценке валидности используются также ***результаты выполнения ранее разработанных тестов***, которые измеряют сходные (*конвергентная валидность*) или различные (*дискриминантная валидность*) для анализируемой методики психологические конструкты. Если новый тест представляет собой сокращенный или упрощенный вариант уже существующего теста, то последний можно с полным основанием считать критериальной мерой. Так, валидизация бланкового теста (типа «бумага-карандаш») может быть осуществлена относительно более сложно организованного и отнимающего много времени теста действия, валидность которого уже установлена. В таких ситуациях новый тест можно считать в лучшем случае

грубой аппроксимацией ранее существующего. Следует отметить, что если новый тест не является более простым или более коротким заменителем ранее доступного теста, то использование последнего в качестве критерия недопустимо.

- **Основные требования к критериям**

Внешний критерий должен отвечать трем основным требованиям:

- он должен быть релевантным (соответствие диагностического инструмента независимому жизненно важному критерию),
- свободным от помех (контаминации),
- и надежным.

Требования свободы от ***контаминации*** вызываются тем, что, например, учебная или производственная успешность зависит от двух переменных: от самого человека, его индивидуальных особенностей, измеряемых методиками, и от ситуации, условий учебы, труда, которые могут привнести помехи, "загрязнить" применяемый критерий. Если преподавателю учебного центра или руководителю цеха станет известно, что данный обучаемый или оператор плохо справился с соответствующим тестом способностей, то это может плохо сказаться на оценке их деятельности. И наоборот, слишком высокие результаты по тесту могли бы подтолкнуть преподавателя или начальника к искусственному завышению академических оценок обучаемого или эффективности оператора соответственно. Такие влияния, очевидно, повышают корреляцию между показателями теста и критерием, искажая действительное положение вещей.

Чтобы предотвратить действие контаминации, ухудшающей критерий валидизации, совершенно необходимо, чтобы лицам, производящим оценку критерия, ничего не было известно о тестовых результатах испытуемого. По этой причине тестовые показатели, используемые при «тестировании теста», должны держаться в строгом секрете. Порой трудно убедить преподавателей, работодателей и других официальных лиц в необходимости такой меры предосторожности. Стремясь использовать всю доступную информацию для принятия практических решений, эти люди могут не понимать того, что показателями теста нельзя пользоваться до тех пор, пока не будут получены критериальные данные и не будет проверена его валидность.

Когда говорят, что критерий должен иметь статистически достоверную надежность, это означает, что он должен отражать постоянство и устойчивость исследуемой функции.

- **Прогностическая валидность**

Прогностическая валидность (*predictive validity*) – это способность теста прогнозировать критериальное событие (например, судить о диагностируемом психологическом свойстве спустя определенное время после измерения). Сведения о прогностической валидности имеют самое непосредственное отношение к раскрытию предсказательной силы методики, выяснению степени обоснованности сформулированного на ее основе ближайшего и более отдаленного прогноза, анализу значимости получаемых в тесте показателей с точки зрения экстраполирования результатов на будущее.

Для определения прогностической валидности разработчик теста должен организовать схему квазиэксперимента по принципу "проспективной валидизации": следует провести тест на обширной выборке испытуемых для изучения корреляции между показателями теста и некоторым критерием, характеризующим измеряемое свойство, но в более позднее время. Многие специалисты по психометрии рассматривают прогностическую валидность как наиболее убедительное подтверждение эффективности теста.

Основная трудность при такой валидизации теста состоит в выборе значимого внешнего критерия. В случае тестов интеллекта кажется разумным, исходя из нашего понятия об интеллекте, использовать будущие успехи в обучении или даже зарабатываемые деньги. Однако, поскольку очевидно существуют и другие переменные, помимо интеллекта, которые связаны с этими критериями, такие как настойчивость, умение ладить с людьми, а также ряд других случайных факторов: хорошее преподавание и наличие вакансий на работе в подходящее время, — можно ожидать, что корреляция с показателями теста интеллекта будет умеренной. Более того, интеллект, возможно, — наиболее простая переменная, для которой может быть спланировано изучение прогностической валидности.

Обычно для тестов способностей и интересов прогностическую валидность продемонстрировать легче, чем для личностных тестов. П. Клайн приводит следующие правила, которым желательно следовать при определении прогностической валидности:

(1) Убедитесь, что выборка испытуемых отражает ту категорию лиц (популяцию), для которой данный тест предназначен, особенно по отношению к полу, возрасту, уровню образования и социальному положению.

(2) Убедитесь, что выборки достаточно велики для получения статистически значимых корреляций, могущих быть затем использованными в факторном анализе. Минимальное количество испытуемых — 200. Размер выборки очень существенен. Если используются множественные корреляции с акцентированием внимания на весовых коэффициентах β (индексе значимости данного теста в прогнозе по данному критерию), то выборка должна быть расщеплена или подвергнута процедуре кросс-валидизации, так как значения весовых коэффициентов β могут изменяться от исследования к исследованию.

(3) При использовании факторного анализа должна быть получена *простая структура*[4].

(4) Должна быть показана надежность используемого критерия. Последняя процедура особенно важна, поскольку неудовлетворительная надежность измерений снижает корреляции.

Для использования факторного анализа пунктов теста при критериальной валидизации в исходный прямоугольный массив ответов в качестве M+1-го столбца в дополнение к M пунктам теста добавляется переменная-критерий (или несколько критериев). В этом случае факторный анализ позволяет одновременно увидеть и то, с какими факторами связан критерий, и то, из каких пунктов состоят эти факторы в данном случае.

Множественный регрессионный анализ предоставляет еще одну удобную возможность для проведения отбора оптимального компактного перечня пунктов тест-опросника: значимые регрессионные веса получают только те пункты, которые вносят свой собственный незаменимый вклад в предсказание критерия. В сочетании с факторным анализом в указанной выше модификации, регрессионный анализ одновременно обеспечивает аналитика пониманием тех психологических факторов (механизмов), благодаря которым достигается (или потенциально может быть достигнуто) предсказание.

В общем виде, коэффициент критериальной валидности есть корреляция между показателями теста и критериальной мерой. Этот коэффициент позволяет охарактеризовать

[4] ***Простая структура*** - это факторное решение, при котором каждый фактор имеет небольшое количество высоких нагрузок, тогда как все остальные нагрузки настолько близки к нулю, насколько возможно. Для достижения простой структуры используются различные процедуры вращения факторной системы координат.

валидность единственным показателем, и поэтому его часто приводят в руководствах к тестам, сообщая его значение для каждого из использованных критериев. Но при прогностической валидизации полученные данные можно дополнительно представить в форме прогностических таблиц или карт прогноза. В прогностической таблице приводятся вероятности различных критериальных исходов испытуемых (например, успехи в обучении), в зависимости от полученного ими результата по тесту. Допустим, в качестве предиктора мы имеем показатель теста на числовые рассуждения, а в качестве прогностического критерия – итоговые оценки учащихся по курсу математики: *A, B, C, D* и *F*, тогда прогностическую таблицу можно представить в следующем виде (см. табл. 15).

Тестовый показатель	**Число случаев**	**Процент получивших критериальную оценку**			
		***D* и ниже**	***C***	***B***	***A***
30 и выше	22	5	0	36	59
20-29	104	9	21	43	27
10-19	71	36	37	24	3
Ниже 10	14	43	36	14	7

Таблица 15. Прогностическая таблица соотношения результатов теста числовых рассуждений и итоговых оценок по курсу математики для 211 учащихся [2].

Во многих практических ситуациях предпочтение может отдаваться дихотомическим критериям в виде «успеха» или «неудачи» в работе, в прохождении учебного курса и т. д. В этих условиях можно построить ***карту прогноза*** (или ***диаграмму ожидаемого отсева***) показывающую вероятность успеха или неудачи для каждого интервала группирования тестовых показателей. Рисунок 11 дает пример такой диаграммы. Базирующаяся на батарее отбора летчиков, разработанной ВВС США, эта диаграмма ожидаемого отсева показывает для каждого значения шкалы результатов теста (станайн – нормированное значение по тесту) процент курсантов, не справившихся с начальным курсом летной подготовки.

На основе данной диаграммы ожидаемого отсева можно предсказать, например, что приблизительно 40% курсантов с тестовым показателем, равным 4 станайнам, потерпят неудачу и приблизительно 60% из них удовлетворительно завершат начальный курс летной подготовки. Аналогичные прогнозы по каждому станайну можно строить и относительно вероятности успеха или неудачи отдельных курсантов. Так, получив тестовый показатель,

равный 4 станайнам, курсант имеет 60 шансов против 40, т. е. 3 шанса против 2, успешно закончить начальный курс летной подготовки.

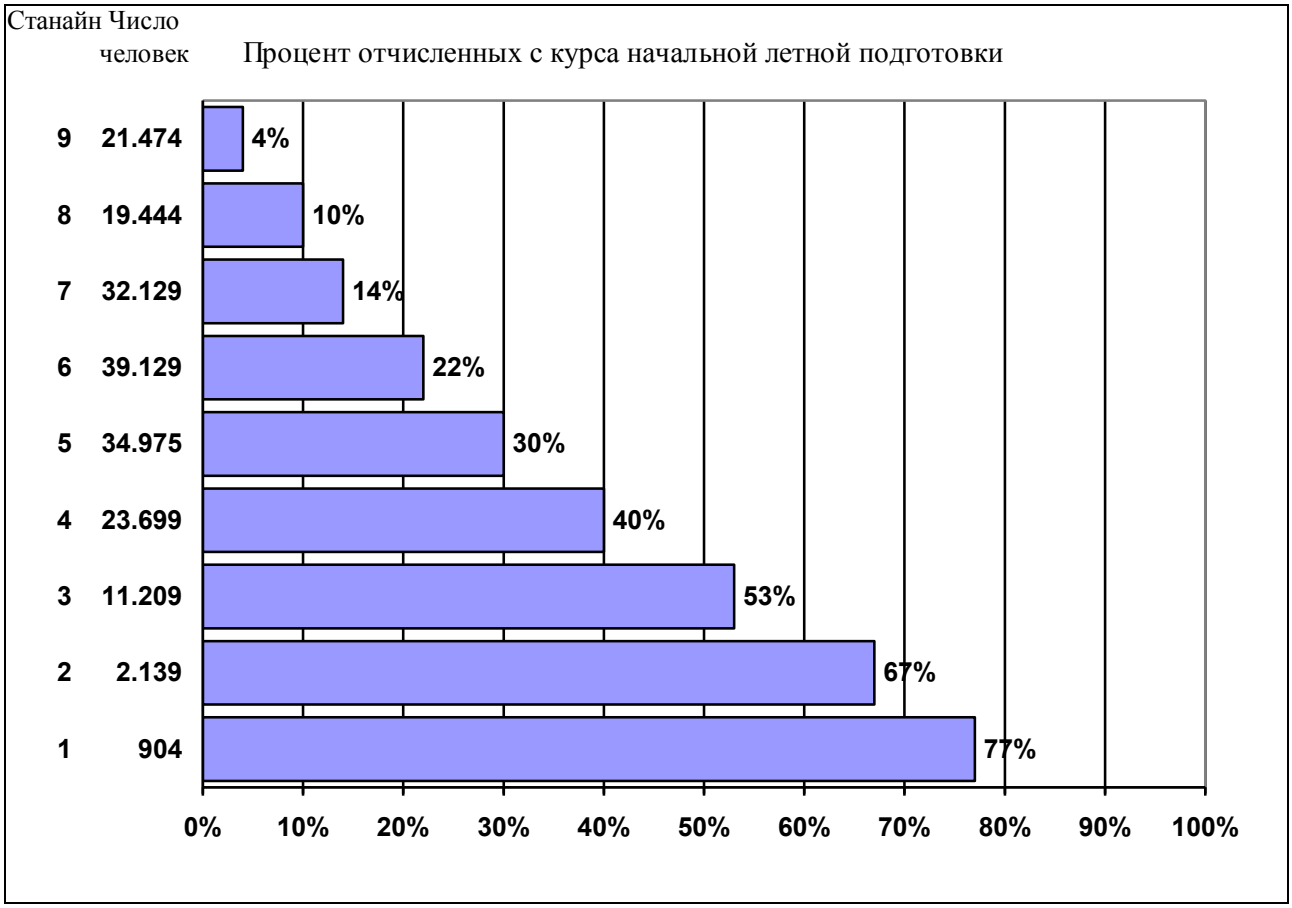


Рисунок 11. Карта прогноза, показывающая связь между выполнением заданий батареи отбора летчиков и отчислением с начального курса летной подготовки [2].

- **Текущая валидность**

Текущая валидность (*concurrent validity*) - это соответствие результатов валидизируемого теста независимому критерию, отражающему состояние исследуемого тестом психологического конструкта в момент проведения исследования. Основной процедурой определения текущей валидности (другое название - ***статусная валидность***) является корреляционный анализ связи результатов теста с критериальными характеристиками исследуемого свойства. Как и для прогностической валидности можно также использовать факторный анализ и множественный регрессионный анализ. Отличие текущей валидности от прогностической валидизации заключается только в том, что оба источника информации об испытуемом - и тест, и критерий - "работают" фактически на одном и том отрезке времени, то есть совпадают в реальном масштабе времени. Например, одним из способов оценки текущей валидности, скажем, теста оперативного мышления, могло бы быть

определение того, как значения этого теста коррелируют с известными навыками группы операторов, эффективность работы которых была оценена в реальных трудовых условиях.

- **Инкрементная валидность**

Инкрементная валидность (*incremental validity)* - "нарастающая пригодность" теста, отражает практическую ценность методики при проведении отбора. Показатель инкрементной валидности указывает на роль теста в улучшении отбора лиц для реальной деятельности, степень улучшения результативности процедуры отбора по сравнению с традиционной, основанной на анализе объективных сведений, документов, бесед, приеме с испытательным сроком и т.п. Инкрементной валидность рассчитывается в зависимости от значений ***индекса*** или ***коэффициента отбора*** (доли принятых на работу по отношению к числу претендентов), коэффициента валидности теста, ***базового уровня*** или ***базисной нормы*** (доли успешно справляющихся с обязанностями работников, отобранных без использования теста, например, с помощью сведений о предыдущей работе, рекомендательных писем и результатов собеседования). Изменение любого из этих условий может повлиять на коэффициент инкрементной валидности. Например, при уменьшении индекса отбора значение коэффициента инкрементной валидности значительно повышается при условии, что используемый тест обладает высокой валидностью. Определение величины коэффициента инкрементной валидности производится с помощью специальных таблиц. В целях иллюстрации воспользуемся одной из таблиц Тейлора-Расселла (см. табл. 16).

Таблица 16 предназначена для использования с базовым уровнем равным 0.60. Числа на пересечении каждой строки (валидность теста) и столбца (индекс отбора) показывают долю успешных работников, отобранных с помощью тестирования. Разность между любым таким числом и базовым уровнем (0.60) отражает коэффициент инкрементной валидности и показывает прирост правильно отобранных работников за счет использования теста.

Очевидно, если индекс отбора равен 100%, т. е. когда пришлось бы принимать на работу всех претендентов, ни один тест, какой бы высокой ни была его валидность, не улучшил бы качества отбора (коэффициент инкрементной валидности = 0.0). Согласно таблицы 16, при индексе отбора равном 0.95, даже абсолютно валидный тест (валидность = 1.00) повысил бы долю успешных работников только на 3% (с 0.60 до 0.63). Напротив, если из поступающих нужно отобрать только 5%, то тест с коэффициентом валидности, равным всего 0.30, может повысить процент удачно отбираемых работников с 60 до 82%. Таким

образом, инкрементная валидность отражает увеличение прогностической валидности, свойственной данному тесту, при заданном базовом уровне и индексе отбора. Инкрементная валидность показывает вклад теста в отбор лиц, которые в дальнейшем будут удовлетворять минимальным требованиям критериальной деятельности.

Валидность теста	**Индекс отбора**										
	0.05	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**0.95**
0.00	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
0.05	0.64	0.63	0.63	0.62	0.62	0.62	0.61	0.61	0.61	0.60	0.60
0.10	0.68	0.67	0.65	0.64	0.64	0.63	0.63	0.62	0.61	0.61	0.60
0.15	0.71	0.70	0.68	0.67	0.66	0.65	0.64	0.63	0.62	0.61	0.61
0.20	0.75	0.73	0.71	0.69	0.67	0.66	0.65	0.64	0.63	0.62	0.61
0.25	0.78	0.76	0.73	0.71	0.69	0.68	0.66	0.65	0.63	0.62	0.61
0.30	0.82	0.79	0.76	0.73	0.71	0.69	0.68	0.66	0.64	0.62	0.61
0.35	0.85	0.82	0.78	0.75	0.73	0.71	0.69	0.67	0.65	0.63	0.62
0.40	0.88	0.85	0.81	0.78	0.75	0.73	0.70	0.68	0.66	0.63	0.62
0.45	0.90	0.87	0.83	0.80	0.77	0.74	0.72	0.69	0.66	0.64	0.62
0.50	0.93	0.90	0.86	0.82	0.79	0.76	0.73	0.70	0.67	0.64	0.62
0.55	0.95	0.92	0.88	0.84	0.81	0.78	0.75	0.71	0.68	0.64	0.62
0.60	0.96	0.94	0.90	0.87	0.83	0.80	0.76	0.73	0.69	0.65	0.63
0.65	0.98	0.96	0.92	0.89	0.85	0.82	0.78	0.74	0.70	0.65	0.63
0.70	0.99	0.97	0.94	0.91	0.87	0.84	0.80	0.75	0.71	0.66	0.63
0.75	0.99	0.99	0.96	0.93	0.90	0.86	0.81	0.77	0.71	0.66	0.63
0.80	1.00	0.99	0.98	0.95	0.92	0.88	0.83	0.78	0.72	0.66	0.63
0.85	1.00	1.00	0.99	0.97	0.95	0.91	0.86	0.80	0.73	0.66	0.63
0.90	1.00	1.00	1.00	0.99	0.97	0.94	0.88	0.82	0.74	0.67	0.63
0.95	1.00	1.00	1.00	1.00	0.99	0.97	0.92	0.84	75	0.67	0.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.86	0.75	0.67	0.63

Таблица 16. Доля «успешных работников», на которую можно рассчитывать при заданном индексе отбора, валидности используемого теста и базовом уровне 0.60 [3].

Инкрементная валидность, вытекающая из использования теста, зависит не только от индекса отбора, но и от базового уровня. Пусть валидность теста равна 0.60, индекс отбора 40%, а базовый уровень равен 50, 10 и 90%. Согласно таблицам Тейлора-Расселла, процент успешных работников повысился бы, соответственно, до 75, 21 и 99%. Таким образом,

увеличение доли успешных работников, которое можно приписать применению теста (инкрементная валидность), составляет 25% при базовом уровне в 50%, но только 11% и 9% при крайних значениях базового уровня. Из этого следует, что хотя внедрение любого валидного теста повысит точность диагностики или прогноза, улучшение точности будет максимальным лишь тогда, когда базовые уровни близки к 50%. При низких базовых нормах (характерно для клинической психологии с редкими патологическими состояниями), это улучшение может оказаться незначительным. В таких случаях использование теста нельзя будет считать оправданным, учитывая издержки, связанные с его проведением и обработкой результатов. В условиях клиники такие издержки включали бы время квалифицированного персонала, которое иначе можно было бы потратить на лечение дополнительных больных. Какое-то количество ложных положительных диагнозов, т. е. нормальных лиц, ошибочно отнесенных к той или иной патологии, еще более увеличило бы эти общие издержки в клинической ситуации.

В целом можно заключить, что инкрементная валидность является полезными показателем эффективности для тех тестов, которые предстоит использовать в процедурах отбора.

- **Конвергентная и дискриминантная валидность**

Конвергентная валидность (*convergent validity*) - степень соответствия двух тестовых методик, направленных на измерение концептуально-родственных конструктов[5]. Эта валидность оценивается по корреляции результатов данного теста с результатами других тестов. Пример конвергентной валидности - наличие значимой корреляции между тестом на "мотивацию достижения" и тестом на "принятие риска" (смелость).

Другой типичный пример, новый тест на умственное развитие сравнивается с известными тестами, такими как Стэнфорд-Бине или Векслер, валидность которых уже известна и доказана. Если показатели участников при испытании нового теста совпадают с показателями при проведении ранее разработанного теста, то новый тест тоже обладает валидностью. Этот пример наглядно иллюстрирует недостаток определения конвергентной валидности. Если уже существует другой валидный тест, достаточно эффективный, чтобы он мог использоваться, то новый тест, который предстоит валидизировать, может быть в какой-то степени ненужным. В самом деле, это будет так, если только он не обладает некоторой значимой характеристикой, не присущей другим валидным тестам. Например,

[5] В книге П. Клайна "Справочное руководство по конструированию тестов" данная валидность переведена как "конкурентная".

если он будет более коротким, простым в использовании, удобным для обработки, или хотя бы будет нравиться испытуемым, то это вполне бы оправдало разработку нового теста тогда, когда существуют другие тесты для измерения данного критерия. С другой стороны, если нет эффективных тестов для измерения данного свойства или особенности, когда новый тест затрагивает иные свойства или особенности индивидуума, тогда ясно, что изучение конвергентной валидности становится затруднительным[6].

В общем, конвергентная валидность полезна тогда, когда есть неудовлетворительно работающие тесты для измерения некоторых переменных, а новые тесты создаются в попытке улучшить качество измерения. В случаях, подобных этому, при изучении конвергентной валидности можно ожидать значимых, но умеренных корреляций. Конвергентная валидность также полезна для установления факта, чего же не измеряет тест. Тест не должен иметь корреляции с другими тестами, измеряющими совершенно иные переменные (конструкты). В этом случае говорят о ***дискриминантной валидности,*** отражающей степень, в которой тест не измеряет тот конструкт, для измерения которого он не предназначен. О наличии дискриминантной валидности говорит отсутствие значимой статистической корреляции между тестовыми показателями, отражающими концептуально независимые свойства.

Итак, показатель конвергентной валидности получают из корреляций (или факторных нагрузок) с другими тестами, которые предназначены для измерения той же переменной. Для эффективного изучения конвергентной валидности П. Клайн приводит следующие правила:

(1) Убедитесь, что выборка испытуемых отражает ту категорию лиц (популяцию), для которой данный тест предназначен, особенно по отношению к полу, возрасту, уровню образования и социальному положению.

(2) Убедитесь, что выборки достаточно велики для получения статистически значимых корреляций, могущих быть затем использованными в факторном анализе. Минимальное количество испытуемых — 200.

(3) Используйте настолько широкое разнообразие других тестов данной переменной, насколько возможно — чтобы убедиться, что корреляция получена благодаря близости групповых факторов, а не специфических. Например, если вы пытаетесь тестом измерить,

[6] Иногда, когда нет тестов для измерения некоторого свойства, исследователи используют экспертные оценки.

то используйте вербальные и невербальные средства измерения созданные различными авторами.

(4) Если используется факторный анализ, убедитесь, что получена простая структура.

(5) При обсуждении результатов четко объясняйте, какие корреляции и нагрузки факторов можно ожидать. Это позволяет читателю судить о психологическом значении этих результатов.

Исследования конвергентной валидности, удовлетворяющие этим критериям, должны дать недвусмысленное свидетельство валидности, которое не может быть методологически опровергнуто.

Кроме проверки статистических гипотез о корреляции значений оценок одного и другого теста, для определения конвергентной валидности может быть использована проверка статистических гипотез об однородности выборочных распределений оценок, полученных при применении тестов. В частности, могут быть проверены статистические гипотезы относительно однородности дисперсий и математических ожиданий по их оценкам дисперсий и средним. Если выполняется требование о нормальности распределения переменных и равенстве их дисперсий, для обнаружения различий между средними двух выборок можно воспользоваться параметрическим *t*-критерием Стьюдента для независимых выборок. Равенство дисперсий в двух выборках можно проверить с помощью *F*-критерия Фишера или более устойчивых (робастных) критериев Ливиня (*Levene test*) и Брауна-Форсайта (*Brown-Forsythe test*). Непараметрическими альтернативами *t*-критерию Стьюдента являются: критерий серий Вальда-Вольфовица, *U*-критерий Манна-Уитни и двухвыборочный критерий Колмогорова-Смирнова.

Для обеспечения дискриминантной валидности одномерного тест-опросника по отношению к фактору социальной желательности А.Г. Шмелев и В.И. Похилько [19] предлагают использовать факторный анализа по методу главных компонент. В этом случае отдельные пункты опросника служат исходными переменными для анализа. Согласно их исследованиям, при факторизации одномерного опросника практически всегда выделяются два главных компонента: один соответствует измеряемому свойству, другой — социальной желательности ответа; сила второго компонента зависит от диагностической ситуации и уровня подозрительности контингента испытуемых. В этом случае значимые факторные нагрузки по первому главному компоненту присваиваются только тем пунктам, которые

имеют незначимые или взаимно уравновешивающие друг друга нагрузки по фактору социальной желательности.

КОМПЛЕКСНАЯ ВАЛИДИЗАЦИЯ

- **Конструктная валидность**

Конструктная валидность (*construct validity*) теста является одним из основных типов валидности и характеризует степень отражения исследуемого психологического конструкта в результатах теста. В качестве конструкта могут выступать практический или вербальный интеллект, эмоциональная устойчивость, интроверсия, понимание речи, переключаемость внимания и т.п. Конструктная валидность определяет область теоретической структуры психологических явлений, измеряемых тестом.

Чтобы продемонстрировать конструктную валидность теста, необходимо настолько полно, насколько это возможно, описать переменную (конструкт), для измерения которой предназначен тест. Это достигается формулированием гипотез о результатах теста в свете всего того, что известно об этой переменной. Конструктная валидность — это мощный метод демонстрации валидности тестов, для которых установление единственного критерия их обоснованности является затруднительным. Вместо одного результата мы должны учитывать одновременно множество результатов. Таким образом, конструктная валидность включает в себя все подходы к определению валидности, перечисленные выше. Вот почему конструктную валидность признали в качестве основного, базисного понятия валидности, включающего все ее остальные виды и позволяющую наиболее точно определить, что и для какой цели измеряется данным тестом.

Среди конкретных методов определения конструктивной валидности в первую очередь необходимо назвать сопоставление исследуемого теста с другими методиками, содержание которых известно. При анализе конструктивной валидности методики обычно формулируют ряд гипотез о том, как будет коррелировать разрабатываемый тест с широким кругом других тестов, направленных на конструкты, находящиеся в теоретически известной или предполагаемой связи и исследуемыми. При этом конструктивная валидность характеризуется не только связями проверяемого теста с близкородственными

показателями (см. Конвергентная валидность), но и с теми, где, исходя из гипотезы, значимых связей наблюдаться не должно (см. Дискриминантная валидность).

Если в ходе конструктивной валидизации теста исследуется его взаимосвязь не с одной методикой, а батареей тестов, то для этих целей используют факторный анализ. Анализ корреляционной матрицы позволяет идентифицировать главные факторы и охарактеризовать тест исходя из полученной факторной структуры, с учетом веса или нагрузки каждого фактора и корреляции теста с каждым из них. Такую корреляцию иногда приводят как ***факторную валидностъ*** (*factorial validity*) теста. Следует отметить, что факторная валидность по существу представляет собой корреляцию теста со всем тем, что есть общего у группы тестов или других индексов поведения. Анализируемое множество переменных может, разумеется, включать в себя как данные тестов, так и данные иного рода. Субъективные оценки и другие меры критерия, наряду с другими тестами, могут быть использованы для исследования факторной структуры конкретного теста и для определения измеряемых им общих черт.

Факторный анализ также необходим и в тех случаях, когда психолог имеет дело с многомерным (многофакторным) тестом и хочет получить «простую» факторную структуру, т. е. такую, в которой максимальное количество пунктов получит значимые нагрузки только по одному фактору. Кроме факторного анализа для этих целей можно использовать кластерный анализ. Алгоритм иерархической кластеризации по методу среднего расстояния группирует пункты в кластеры (производит разбиение объектов на множество непересекающихся классов[7]), соответствующие, как правило, факторам, получаемым после варимакс-вращения (*Varimax procedure*) главных компонент.

Важным аспектом конструктной валидности является внутренняя согласованность (***внутренняя валидность***), отражающая то, насколько определенные пункты (задания, вопросы) или субтесты, составляющие материал теста, подчинены основному направлению теста как целого, ориентированы на изучение одних и тех же конструктов. В публикуемой информации о некоторых тестах можно встретить утверждение, что валидность теста была установлена методом внутренней согласованности. Существенной особенностью этого метода является использование в качестве критерия валидизации суммарного показателя самого теста.

[7] В настоящее время разработаны методы кластерного анализа, например, "анализ клик" или "клайк-анализ" (от англ. clique - клика), которые дают разбиение объектов на пересекающиеся классы, что делает кластеризацию объектов более точной.

Анализ внутренней согласованности осуществляется путем коррелирования ответов на каждое задание с общим результатом теста. Для этих целей можно, например, вычислить бисериальные коэффициенты корреляции между исходами («справился — не справился») каждого задания и суммарным показателем теста (см. Коэффициент дискриминации). В этом случае сохраняются только те задания, для которых отмечена значимая корреляция с тестом в целом (не менее 0.25). Задания с низкой корреляцией с общим результатом теста должны быть переработаны или исключены. Если тест состоит из заданий, прошедших такого рода отбор, то можно говорить о его внутренней согласованности, поскольку каждое его задание дифференцирует респондентов в том же направлении, что и тест в целом.

Иногда для оценки внутренней согласованности теста приспосабливается метод контрастных групп, которые в этом случае формируются из испытуемых с самыми высокими и с самыми низкими суммарными показателями по данному тесту (примером могут служить выборки В2 и В3, см. рис. 9). Результаты выполнения каждого задания теста группой с верхним значением критерия сравнивается затем с соответствующими результатами группы с нижним значением критерия. Задания, по которым не удалось обнаружить существенно большей доли «правильных» (совпадающих с ключом) ответов в группе с верхним значением критерия по сравнению с группой с низким значением критерия, признаются невалидными и либо отбрасываются, либо перерабатываются.

Очевидно, что корреляции, отражающие внутреннюю согласованность теста, являются по существу мерой его однородности. Поскольку это свойство помогает охарактеризовать область поведения или отдельную черту, выборочно проверяемые тестом, то степень однородности теста имеет отношение к его конструктной валидности. Тем не менее, вклад данных о внутренней согласованности теста в его валидизацию носит ограниченный характер. При отсутствии внешних по отношению к тесту данных, мало что можно узнать о том, что он в действительности измеряет.

Еще один источник данных для валидизации конструкта обеспечивают эксперименты, в которых исследуется влияние выбранных переменных на показатели теста. При проверке валидности теста, предназначенного, например, для использования в программе индивидуализированного обучения, есть только один путь — сравнить показатели тестирования до и после экспериментального обучения. Логическое обоснование такого теста требует низких показателей при первом тестировании, проводимом до

соответствующего обучения, и высоких показателей при втором тестировании, после обучения. То же соотношение может проверяться и для отдельных заданий теста. В идеале с каждым заданием до обучения должно справиться минимальное, а после обучения — максимальное число учеников. Задания, с которыми мало кто справляется в обоих случаях, слишком трудны, а те, с которыми справляются почти все и до и после обучения, слишком доступны с точки зрения целей, преследуемых тестом. Если же многие в первый раз справляются, а во второй раз не справляются с заданием, то что-то неладно или с этим заданием, или с обучением, или с тем и другим.

Другим примером экспериментальной валидизации может служить ситуация, когда тест, предназначенный, например, для измерения склонности к тревоге (*anxiety-proneness*), проверяется через предъявление его испытуемым до и после того, как они были помещены в обстановку, провоцирующую состояние тревоги (например, выполнение сложной и ответственной деятельности, когда любая ошибка может иметь значимые последствия). Исходные тестовые показатели тревожности можно затем соотнести с физиологическими и иными показателями выражения тревоги во время и после экспериментального воздействия. Другую (дифференциальную) гипотезу в отношении теста тревожности можно оценить, проводя тест до и после вызывающего тревогу события и наблюдая за тем, происходит ли существенное увеличение тестовых показателей при втором тестировании. Положительные результаты такого эксперимента будут свидетельствовать о том, что тестовые показатели отражают текущий уровень тревожности. Аналогичным образом можно планировать эксперименты для проверки гипотез относительно любого конкретного психического свойства, измеряемого тестами.

Понятие конструктной валидности можно пояснить на примере, который использовал П. Клайн [7]. Он сформулировал следующие гипотезы, подлежащие проверке при установлении конструктной валидности теста оральных черт личности — The Oral Pessimism Questionnaire (OPQ):

(1) OPQ будет коррелировать положительно, но умеренно (ввиду их невысокой эффективности) с другими тестами, направленными на выявление оральных черт личности.

(2) Из описания синдрома "орального пессимизма" должна наблюдаться умеренная корреляция с нейротизмом.

(3) Поскольку 16-факторный личностный опросник Кэттелла не предназначен для измерения параметров, подобных данному синдрому, то с этим опросником не должно быть никаких корреляций.

(4) Поскольку OPQ является личностным тестом, не должно быть значимых корреляций с переменными способностей или мотивов.

Данные гипотезы иллюстрируют необходимость показать, при исследовании конструктной валидности, чего тест не измеряет (***дискриминантная валидность***), наряду с тем, что он измеряет (***конвергентная валидность***).

Если перечисленные гипотезы получат подтверждение их истинности, все еще остается спорным, что продемонстрирована конструктная валидность теста OPQ как средства измерения совокупности личностных черт, определяемых как "оральный пессимизм". Дальнейший, более непосредственный способ продемонстрировать конструктивную валидность теста может состоять в формулировании пятой гипотезы:

(5) Испытуемые, имеющие высокую выраженность измеряемых черт личности, покажут по OPQ более высокие показатели, чем те, у которых она низкая (***текущая валидность***).

Общую процедуру определения конструктной валидности П. Клайн представил следующим образом:

(1) Перечислите точно гипотезы, касающиеся переменных, с которыми данный тест должен коррелировать (***конвергентная валидность***).

(2) Перечислите точно гипотезы, касающиеся переменных, с которыми данный тест не должен коррелировать (***дискриминантная валидность***).

(3) Укажите группы, которые должны давать низкие и высокие показатели по данному тесту (***текущая валидность***).

(4) Сформулируйте гипотезу о месте данного теста в факторном пространстве. Эта гипотеза подобна гипотезам из выше приведенных пунктов (1) и (2).

Эти четыре гипотезы должны затем быть проверены на больших выборках, соответствующим образом сформированных, как указано в процедурах для установления конкурентной валидности. Специфические группы должны быть достаточно большими, не только для выявления статистически значимых различий, но также такими, чтобы с уверенностью могли быть сделаны обобщения.

ТРЕБОВАНИЯ К ВЫБОРКЕ ИСПЫТУЕМЫХ ПРИ ИЗУЧЕНИИ ВАЛИДНОСТИ

Выборки работников предприятий, доступные исследователям при валидизации тестов, обычно слишком малы, чтобы дать устойчивую оценку корреляции между прогнозирующим показателем и критерием. По той же причине получаемые коэффициенты могут оказаться слишком низкими, чтобы достичь статистической значимости в используемой для валидизации выборке, и потому не пригодными в качестве доказательства валидности теста. Примерно половина выборок работников промышленных предприятий, используемых в исследованиях валидности, включает не более 40÷50 человек. При таких малых выборках, согласно А. Анастази и С. Урбина, валидизация через предсказание критерия технически не осуществима [3]. Для проведения критериальной валидизации П. Клайн предлагает использовать 200 испытуемых как минимальное количество для получения статистически значимых корреляций, могущих быть затем использованных в факторном анализе [7]. Согласно И.М. Кондакова [9], для проверки дискриминативности, надежности и валидности опросников объем выборки обычно не превышает 50÷100 чел.

Как и в случае с надежностью, важно точно определить характер группы, на которой вычисляется коэффициент валидности теста. Один и тот же тест может измерять различные функции, если его дать лицам разного возраста, пола, уровня образования, рода занятий и т. д. Люди с разным жизненным, учебным и профессиональным опытом могут, например, воспользоваться разными методами для решения одной и той же тестовой задачи. Следовательно, тест может обладать высокой валидностью относительно заданного критерия в одной популяции и низкой или нулевой валидностью — в другой. Поэтому необходимо убедиться, что выборка испытуемых отражает ту категорию лиц (популяцию), для которой данный тест предназначен, особенно по отношению к полу, возрасту, уровню образования и социальному положению.

Интересы и мотивация могут также оказывать существенное влияние на валидность тестов. Так, если кандидатам на рабочие места эта работа мало интересна, они, вероятно, будут выполнять ее без особого усердия, какими бы ни были их показатели по соответствующим тестам способностей. Для таких лиц корреляция между результатами теста способностей и качеством выполнения работы будет низкой, тогда как для заинтересованных и высоко мотивированных такая корреляция может оказаться весьма значительной.

Вопрос разнородности выборки имеет для измерения валидности такое же значение, как и для измерения надежности (см. Требования к выборке испытуемых при изучении надежности), поскольку обе характеристики обычно приводятся в виде коэффициентов корреляции. Напомним, что при прочих равных условиях, чем шире размах распределения показателей, тем выше будет корреляция. Это обстоятельство необходимо иметь в виду при интерпретации коэффициентов валидности, приводимых в руководствах к тестам.

Специфическая проблема, присущая многим выборкам валидизации, связана с *предварительным отбором (preselection)* испытуемых. Например, новый тест, валидизируемый для целей профотбора, может проводиться на группе недавно нанятых работников, в отношении которых со временем будут доступны такие меры критерия, как эффективность труда. Вполне вероятно, однако, что эти работники представляют собой верхнюю (лучшую) часть выборки из всех тех, кто хотел поступить на эту работу. Поэтому нижний конец распределения тестовых показателей и критериальных мер в такой выборке окажется обрезанным. Эффектом такого предотбора, естественно, будет снижение коэффициента валидности. При последующем использовании теста, когда его будут проводить со всеми поступающими на работу в целях их отбора, можно ожидать некоторого повышения его валидности.

Повышение требований при приеме на работу также может приводить к снижению коэффициентов валидности в связи с возрастанием однородности группы принятых на работу кандидатов. В этом случае можно наблюдать падение корреляции между результатами тестов и критериальными показателями деятельности, которое не свидетельствует о том, что прогнозирующие показатели стали менее валидными.

ОБЩИЕ ПРИНЦИПЫ ДЛЯ ИНТЕРПРЕТАЦИИ КОЭФФИЦИЕНТОВ ВАЛИДНОСТИ

- **Форма связи теста и критерия**

Для правильной интерпретации коэффициента валидности следует принимать во внимание и *форму связи* между тестом и критерием. Вычисление пирсоновского коэффициента корреляции предполагает, что эта связь линейна и остается неизменной во всем диапазоне распределения. Однако, исследование связи тестовых показателей с выполнением работы

показало, что эти условия в ряде случаев не выполняются. Пусть для выполнения некоторой работы требуется лишь минимальный уровень понимания читаемого, достаточный для прочтения инструкций, названий и т. д. Но как только этот минимальный уровень превзойден, то от дальнейшего развития данного умения успешность выполнения работы уже не зависит, т. е. между тестом и выполнением работы существуют нелинейные отношения. Изучение двумерного распределения или диаграммы рассеяния, построенной по показателям теста на понимание читаемого и мерам критерия, в этом случае показало бы, что уровень выполнения работы растет, пока умение понимать читаемое не достигает требуемой степени, после чего он остается примерно тем же. Следовательно, точки на диаграмме группируются вокруг кривой, а не прямой линии.

В других случаях линия наилучшего соответствия может быть и прямой, но точки, соответствующие индивидуальным данным, могут отклоняться от нее в верхнем конце шкалы больше, чем в нижнем. Предположим, что успешное выполнение теста академических способностей — необходимое, но не достаточное условие для успешного завершения некоторого учебного курса. Это значит, что учащиеся с низкими показателями по данному тесту получат скорее всего неудовлетворительные оценки, тогда как среди учащихся с высокими показателями одни получат положительные оценки, а другие, из-за недостаточной мотивации, отсутствия интереса или других неблагоприятных условий, не сдадут экзамена. В этой ситуации будет наблюдаться большая вариативность выполнения критериальной деятельности у учащихся с высокими тестовыми показателями, чем с низкими. Такое условие в двумерном распределении называется гетероскедастичностью. Определение корреляции по Пирсону предполагает наличие гомоскедастичности, т. е. одинаковую дисперсию (вариабельность) критерия по всей области двумерного распределения. В приведенном примере двумерное распределение было бы веерообразным — широким в верхнем конце и узким в нижнем. Уже визуального анализа двумерного распределения обычно бывает достаточно для установления характера связи между тестом и критерием. Прогностические таблицы и диаграммы ожидаемого отсева также достаточно хорошо выявляют относительную эффективность теста на разных уровнях.

Нет единого ответа на вопрос, какова должна быть величина коэффициента валидности, так как при интерпретации коэффициента валидности нужно учитывать ряд сопутствующих обстоятельств. Разумеется, корреляция должна быть достаточно высокой для того, чтобы быть статистически значимой на приемлемом уровне, таком как 0.01 или 0.05. Иными словами, прежде чем делать какие-либо выводы о валидности теста, нужно

иметь обоснованную уверенность в том, что полученный коэффициент валидности не появился в результате случайных колебаний выборки из генеральной совокупности с нулевой корреляцией.

- **Стандартная ошибка оценки**

Установив значимую корреляцию между тестовыми показателями и критерием, необходимо еще оценить ее величину в аспекте тех целей, ради которых и создавался данный тест. Если мы собираемся предсказывать точное значение критериального показателя у конкретных лиц (скажем, средний балл обучаемого), коэффициент валидности можно интерпретировать исходя из ***стандартной ошибки оценки*** (*Standard Error of Estimate* - ***SEE***), которая аналогична стандартной ошибке измерения, обсуждавшейся в связи с надежностью (см. Общие принципы для интерпретации коэффициентов надежности). Напомним, что стандартная ошибка измерения указывает допустимый предел возможной ошибки индивидуального показателя в результате ненадежности теста. Аналогично этому, ошибка оценки указывает допустимый предел возможной ошибки прогнозируемой величины индивидуального критериального показателя в результате недостаточной валидности теста.

Ошибка оценки вычисляется по следующей формуле:

$$SEE = SD_y \cdot \sqrt{1 - r_{xy}^2}$$

где SD_y - стандартное отклонение критериального показателя; r^2_{xy} – квадрат коэффициента валидности.

Заметим, что при полной валидности (r^2_{xy} = 1.00) ошибка оценки была бы равна нулю. С другой стороны, если валидность теста равна нулю, то ошибка оценки достигает величины стандартного отклонения распределения критерия (SD_y). При этих условиях вероятность правильного прогноза не превышает вероятности случайного угадывания, и диапазон ошибки предсказания равен ширине распределения критериальных показателей. Между этими двумя пределами и будут заключаться ошибки оценки, соответствующие тестам с варьирующей валидностью.

Из формулы *SEE* видно, что величина $\sqrt{1 - r_{xy}^2}$ указывает на величину ошибки относительно ошибки простого угадывания, т.е. при нулевой валидности. Иными словами, если $\sqrt{1 - r_{xy}^2}$ = 1.00, то ошибка оценки столь же велика, как и при случайном угадывании

критериального показателя у конкретного испытуемого. Использование такого теста не дало бы нам никакого выигрыша в точности предсказания. Если же коэффициент валидности равен 0.80, то $\sqrt{1-r_{xy}^2}$ = 0.60, и максимальная ошибка составляет 60% от величины той, которая была бы при случайном угадывании. Это означает, что тест позволяет делать прогнозы о критериальном выполнении индивида с ошибкой на 40% меньше, чем в случае угадывания.

Может показаться, что даже при такой необычайно высокой валидности, как 0.80, ошибка предсказываемых показателей довольно значительна. Если бы главной функцией психологических тестов было предсказание точного положения индивидуума в критериальном распределении, такая перспектива выглядела бы совершенно обескураживающей. Когда мы рассматриваем тесты в аспекте ошибки оценки, большинство из них представляются не особенно эффективными. Однако чаще всего при тестировании нет необходимости предсказывать точный результат критериальной деятельности каждого обследуемого человека, но требуется лишь определить, кто из них превзойдет некоторый минимальный стандарт выполнения, или критический показатель выбранной в качестве критерия деятельности: каковы шансы субъекта закончить курс специальной подготовки, преуспеть в качестве руководителя, кто из обследуемых скорее всего будет хорошим оператором. Такая информация полезна не только для профотбора, но и для профориентации.

Как правило, один тест никогда не может полностью предсказать, например, эффективность выполнения деятельности, так как слишком многочисленные и различные факторы влияют на прогнозируемый критерий. Поэтому коэффициенты критериальной валидности, в отличие от коэффициентов надежности, редко превышают r = 0.40. Для упрощенной интерпретации полученных значений критериальных коэффициентов валидности можно воспользоваться таблицей 17.

По существу, тест может заметно повысить свою предсказуемостную эффективность, если для него будет установлена любая значимая корреляция с критерием, какой бы низкой она ни была. При некоторых обстоятельствах валидность порядка 0.20÷0.30 уже оправдывает включение теста в программу отбора. Для многих целей тестирования оценивание тестов с точки зрения их стандартной ошибки оценки является неоправданно строгим. В большинстве случаев должны применяться другие способы оценивания тестов, те, которые

бы учитывали типы решений, принимаемых на основе их результатов. Одним из примеров может служить инкрементная валидность, которая учитывает не только исходный коэффициент валидности теста, но также коэффициент отбора (долю принятых на работу по отношению к числу претендентов и базовый уровень (долю успешно справляющихся с обязанностями работников, отобранных без использования теста). Подробно этот вопрос был рассмотрен выше (см. Инкрементная валидность).

Величина коэффициента валидности	Интерпретация
Выше 0.35	Очень хорошая.
0.21 ÷ 0.35	Можно использовать.
0.11 ÷ 0.20	Можно использовать в зависимости от обстоятельств.
ниже 0.11	Тест не следует использовать.

Таблица 17. Общие принципы для интерпретации коэффициентов валидности [10].

- **Коррекция валидности от значений надежности теста и критерия**

Валидность теста всегда ограничена его надежностью. Часть тестового балла, приходящаяся на случайную ошибку, не коррелирует с критерием. Поэтому, если надежность теста меньше 1, то есть истинный балл не совпадает с тестовым, то корреляция между тестом и критерием будет занижена. Если нам известна их надежность, то мы можем откорректировать занижение корреляции:

$$r_{xy,cor} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}$$

где $r_{xy,cor}$ – откорректированный коэффициент корреляции между тестом X и критерием Y (значение валидности теста с учетом надежности), r_{xy} – неоткорректированный коэффициент корреляции (исходная критериальная валидность теста), r_{xx} и r_{yy} – надежность теста X и критерия Y, соответственно.

- **Простая стратегия отбора**

Простую стратегию отбора претендентов иллюстрирует рисунок 12. На этом рисунке показан рост количества успехов вследствие использования отборочного теста.

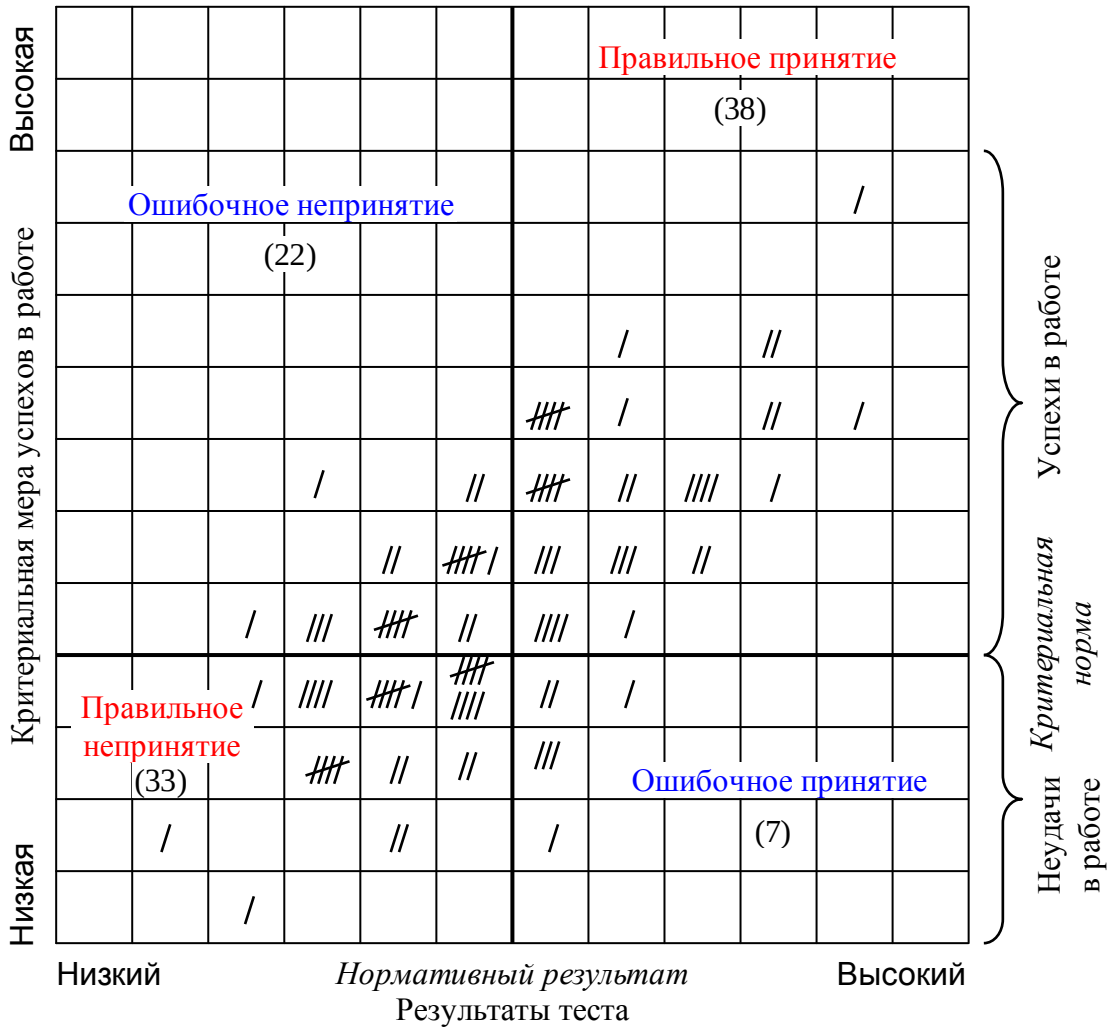


Рисунок 12. Рост количества успехов вследствие использования отборочного теста [2].

Предположим, 100 человек, поступающих на работу, выполнили тест способностей и по прошествии какого-то времени были оценены их успехи в выполнении своих обязанностей. На рисунке 12 изображено соответствующее двумерное распределение тестовых показателей и мер успешного выполнения работы. Корреляция между этими двумя переменными чуть ниже 0.70 (*прогностическая валидность*). Необходимый минимум выполнения работы, или критический показатель, отмечен на диаграмме жирной горизонтальной линией. 40 случаев, лежащих ниже этой линии, соответствуют числу людей, не справившихся с работой, а 60 случаев выше нее - числу успешно работающих. Если на работу принимаются все 100 человек, подавших заявление, то, следовательно, 60% справятся с ней (базовый уровень). Подобным же образом, если бы меньшее число работников нанималось наугад, безотносительно к результатам тестирования, доля успешно справившихся с работой была бы, вероятно, близка к 60%. Предположим, однако, что тестовые показатели используют для отбора из 100 претендентов 45 наиболее

перспективных работников (индекс отбора = 0.45). В таком случае были бы выбраны 45 человек, попадающие в область справа от жирной вертикальной линии. На диаграмме видно, что из этих 45 человек **7** попадают ниже жирной горизонтальной линии, т. е. в разряд несправившихся с работой, и составляют долю ошибочно принятых, а 38 человек - в разряд успешных работников. Процент успешно справившихся с работой теперь уже равен не 60%, а 84% (т. е. 38/45 = 0.84). Это увеличение на 24% обусловлено применением теста в качестве инструмента отбора (*инкрементная валидность*). Заметим, что ошибками показателя предсказываемого критерия, не влияющими на принятие решение, можно пренебречь. Селективную эффективность теста будут снижать только те ошибки предсказания, которые ведут к пересечению линии критического показателя и, следовательно, к помещению индивидуума в ошибочную категорию.

Для полной оценки эффективности теста как инструмента отбора необходимо также изучить другую категорию случаев, отображенную на рисунке 12. Это категория ошибочно непринятых, включающая 22 человека, у которых показатели по тесту ниже критического уровня, а показатели критериальной деятельности выше такового. Исходя из полученных данных, можно приблизительно оценить, что 22% всей выборки претендентов на получение работы, являясь потенциально успешными работниками, будут потеряны в том случае, когда данный тест применяется в качестве инструмента отбора с выбранным таким образом критическим показателем. Устанавливая уровень критического показателя по тесту, следует учитывать процент случаев ошибочного отказа в приеме, а также процент успешных и неуспешных работников в группе отобранных. В определенных ситуациях уровень устанавливаемого критического показателя должен быть достаточно высоким, чтобы почти полностью исключить возможные неудачи. Это необходимо, когда характер работы таков, что недостаточно квалифицированный работник может нанести серьезный ущерб или вред. В качестве примера здесь уместно указать на отбор пилотов гражданской авиации или операторов АЭС. При других обстоятельствах бывает важнее нанять как можно больше квалифицированных работников, идя на риск принять и больше неспособных к данному роду деятельности. В последнем случае число ошибочных отказов сокращается за счет выбора более низкого уровня критического показателя. К другим факторам, которые обычно влияют на уровень критического показателя, относятся число претендентов, количество вакансий и сроки, в которые эти вакансии необходимо заполнить.

Во многих кадровых решениях коэффициент отбора определяется практическими требованиями конкретной ситуации. В одних случаях соотношение спроса и предложения обусловливает, например, прием 40%, а в других — 75% претендентов (с лучшими показателями, разумеется). Если коэффициент отбора не диктуется внешними обстоятельствами, то критический показатель по тесту может устанавливаться на уровне, обеспечивающем наилучшую дифференциацию двух групп по критериальной деятельности. Приблизительно это можно сделать, сравнивая распределение показателей теста в группах «успешных» и «неуспешных» работников.

На языке теории принятия решений, представленный на рисунке 12 пример иллюстрирует простую стратегию отбора претендентов. В более широком смысле, стратегия — это способ использования информации для выработки решения в отношении определенного круга лиц. В данном случае стратегия состоит в приеме 45 человек с самыми высокими тестовыми показателями. Увеличение доли успешно справляющихся со своей работой лиц с 60 до 84% могло бы послужить основанием для оценивания чистой выгоды от использования теста.

- **Отношение валидности к продуктивности.**

Во многих практических ситуациях требуется оценить эффективность теста для профотбора не по проценту лиц, преодолевших «планку» минимальных требований к деятельности, а по предельной продуктивности труда отобранных с его помощью работников. Как реальный уровень квалификации работников (или выполнения ими критериальной деятельности), нанятых по результатам тестирования, сравнить с уровнем общей выборки кандидатов, которые могли бы быть приняты на работу без проведения данного теста? Брогден (Brogden, 1946b) первым показал, что ожидаемый прирост продуктивности прямо пропорционален валидности теста. Так, улучшение от применения теста с валидностью 0.50 составляет 50% улучшения, ожидаемого при использовании абсолютно валидного теста.

Связь между валидностью теста и ожидаемым повышением критериальных достижений видна из таблицы 18. Выражая критериальные показатели в виде стандартных показателей со средним равным нулю и *SD* = 1, эта таблица содержит ожидаемые средние критериальных показателей работников, отобранных при заданном коэффициенте отбора с помощью теста, имеющего определенную валидность. В этом контексте средняя базисная продуктивность, соответствующая деятельности работников, набранных без использования

теста, приводится в колонке нулевой валидности. Использовать тест с нулевой валидностью — это все равно, что не использовать никаких тестов. Покажем, как пользоваться этой таблицей. Предположим, приему подлежат 20% претендентов с самыми высокими показателями (коэффициент отбора 0.20), причем отбор производится с помощью теста, валидность которого равна 0.50. По таблице 18 находим, что средний критериальный показатель в отобранной группе превышает средний показатель базисной продуктивности на 0.7 *SD*. При том же коэффициенте отбора (0.20) и применении идеального теста (с коэффициентом валидности 1.00) средний критериальный показатель принятых на работу претендентов составил бы уже 1.40, т. е. оказался бы ровно в два раза выше, чем при использовании теста с валидностью 0.50. Подобная прямая линейная зависимость имеет место в пределах любой строки таблицы 18. Например, при коэффициенте отбора 0.60 тест с валидностью 0.25 дает средний критериальный показатель 0.16, в то время как тест с валидностью 0.50 обеспечивает средний критериальный показатель 0.32. Опять-таки удвоение валидности ведет к удвоению показателя продуктивности.

Интересен анализ продуктивности в связи с валидностью тестов, используемых для отбора кадров, выполненный Шмидтом и его коллегами [3]. Выбрав в качестве иллюстративного образца работу программиста в федеральном правительстве, эти исследователи оценили в долларовом эквиваленте повышение продуктивности в результате использования в течение года теста компьютерных способностей (*computer aptitude test*) (коэффициент валидности равен 0.76) при отборе наемных работников. Ожидаемая прибыль рассчитывалась для девяти индексов отбора в диапазоне от 0.05 до 0.80, и для пяти коэффициентов валидности методик предварительного отбора (до использования теста компьютерных способностей) — от нуля (случайный отбор) до 0.50.

Результаты показали впечатляющий прирост продуктивности труда от использования теста при всех этих условиях. Когда отбор на основе теста сравнили со случайным отбором (валидность предварительного отбора 0.00), прирост производительности в долларовом эквиваленте колебался от $97.2 млн. при индексе отбора 0.05 до $16.5 млн. при индексе отбора 0.80. При валидности предварительного отбора 0.50 соответствующий прирост колебался от $33.3 млн. до $5.6 млн. Оценки основывались на предположении, что отбор начинается с претендентов, имеющих высшие показатели по тесту, и продолжается до тех пор, пока не будет достигнуто заданное значение индекса отбора.

Индекс отбора	Коэффициент валидности																				
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**	**1.00**
0.05	0.00	0.10	0.21	0.31	0.42	0.52	0.62	0.73	0.83	0.94	1.04	1.14	1.25	1.35	1.46	1.56	1.66	1.77	1.87	1.98	2.08
0.10	0.00	0.09	0.18	0.26	0.35	0.44	0.53	0.62	0.70	0.79	0.88	0.97	1.05	1.14	1.23	1.32	1.41	1.49	1.58	1.67	1.76
0.15	0.00	0.08	0.15	0.23	0.31	0.39	0.46	0.54	0.62	0.70	0.77	0.85	0.93	1.01	1.08	1.16	1.24	1.32	1.39	1.47	1.55
0.20	0.00	0.07	0.14	0.21	0.28	0.35	0.42	0.49	0.56	0.63	0.70	0.77	0.84	0.91	0.98	1.05	1.12	1.19	1.26	1.33	1.40
0.25	0.00	0.06	0.13	0.19	0.25	0.32	0.38	0.44	0.51	0.57	0.63	0.70	0.76	0.82	0.89	0.95	1.01	1.08	1.14	1.20	1.27
0.30	0.00	0.06	0.12	0.17	0.23	0.29	0.35	0.40	0.46	0.52	0.58	0.64	0.69	0.75	0.81	0.87	0.92	0.98	1.04	1.10	1.16
0.35	0.00	0.05	0.11	0.16	0.21	0.26	0.32	0.37	0.42	0.48	0.53	0.58	0.63	0.69	0.74	0.79	0.84	0.90	0.95	1.00	1.06
0.40	0.00	0.05	0.10	0.15	0.19	0.24	0.29	0.34	0.39	0.44	0.48	0.53	0.58	0.63	0.68	0.73	0.77	0.82	0.87	0.92	0.97
0.45	0.00	0.04	0.09	0.13	0.18	0.22	0.26	0.31	0.35	0.40	0.44	0.48	0.53	0.57	0.62	0.66	0.70	0.75	0.79	0.84	0.88
0.50	0.00	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36	0.40	0.44	0.48	0.52	0.56	0.60	0.64	0.68	0.72	0.76	0.80
0.55	0.00	0.04	0.07	0.11	0.14	0.18	0.22	0.25	0.29	0.32	0.36	0.40	0.43	0.47	0.50	0.54	0.58	0.61	0.65	0.68	0.72
0.60	0.00	0.03	0.06	0.10	0.13	0.16	0.19	0.23	0.26	0.29	0.32	0.35	0.39	0.42	0.45	0.48	0.52	0.55	0.58	0.61	0.64
0.65	0.00	0.03	0.06	0.09	0.11	0.14	0.17	0.20	0.23	0.26	0.28	0.31	0.34	0.37	0.40	0.43	0.46	0.48	0.51	0.54	0.57
0.70	0.00	0.02	0.05	0.07	0.10	0.12	0.15	0.17	0.20	0.22	0.25	0.27	0.30	0.32	0.35	0.37	0.40	0.42	0.45	0.47	0.50
0.75	0.00	0.02	0.04	0.06	0.08	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.30	0.32	0.33	0.36	0.38	0.40	0.42
0.80	0.00	0.02	0.04	0.05	0.07	0.09	0.11	0.12	0.14	0.16	0.18	0.19	0.21	0.22	0.25	0.26	0.28	0.30	0.32	0.33	0.35
0.85	0.00	0.01	0.03	0.04	0.05	0.07	0.08	0.10	0.11	0.12	0.14	0.15	0.16	0.18	0.19	0.20	0.22	0.23	0.25	0.26	0.27
0.90	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
0.95	0.00	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.07	0.07	0.08	0.08	0.09	0.09	0.10	0.10	0.11

Таблица 18. Средние стандартных критериальных показателей принятых на работу в зависимости от валидности теста и индекса отбора [3].

- **Понятие полезности в теории принятия решений**

Именно теория принятия решений позволяет оценить тесты по их эффективности в конкретной ситуации. Такая оценка учитывает не только валидность теста при предсказании определенного критерия, но и ряд других параметров, включая базовый уровень и индекс отбора. Еще одним важным параметром является *относительная полезность* (*utility*) ожидаемых результатов: определенным образом оцененная благоприятность или неблагоприятность каждого из них.

Следует отметить, что требуемые в моделях принятия решений оценки имеют отношение не к абсолютной, а лишь к относительной ценности различных результатов. При выборе стратегии решения цель заключается в максимизации ожидаемой полезности на всем множестве результатов. Схема простой стратегии, представленная на рисунке 13, поможет прояснить суть метода. На этой схеме изображена стратегия принятия решений в ситуации, отображенной на рисунке 12, когда в группе претендентов на получение работы проводился всего один тест и на основе сравнения индивидуальных показателей с критическим показателем по этому тесту выносились решения о приеме на работу или отказе. В этой ситуации имеется всего четыре возможных исхода, или результата: правильное/ошибочное принятие и правильное/ошибочное непринятие. Вероятность каждого результата можно вычислить, исходя из числа претендентов, попадающих в каждый квадрант на рисунке 12.

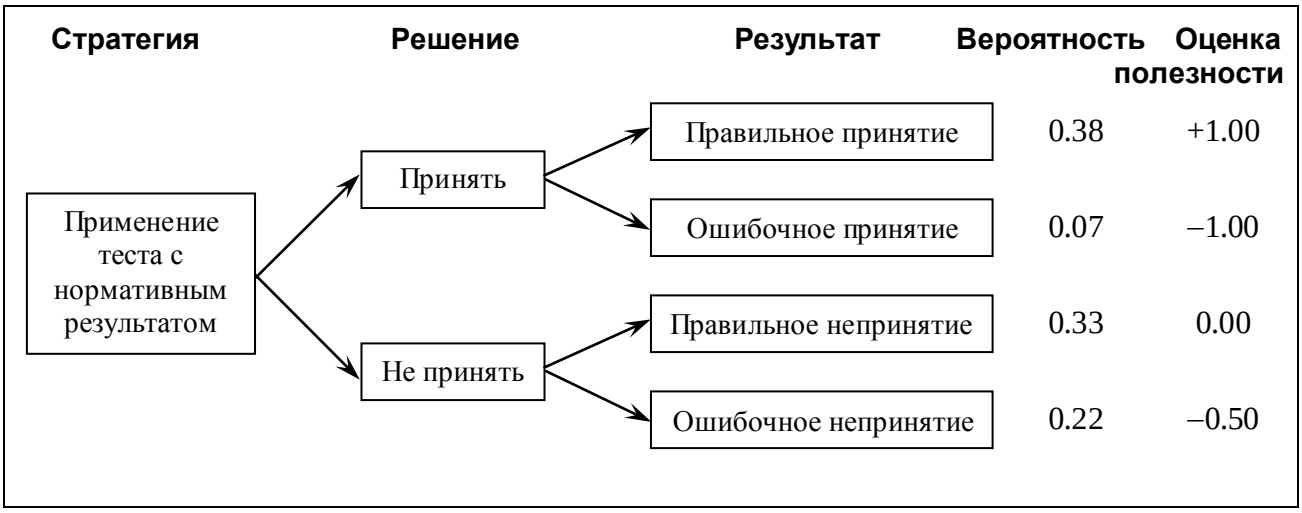


Рисунок 13. Простая стратегия принятия решения и оценка полезности [3].

Кроме того, нужно знать полезности различных результатов, выраженные в единой шкале. Эти гипотетические величины, полученные с помощью любой оценочной процедуры, приведены в последнем столбце на рисунке 13. Общую ожидаемую полезность стратегии можно найти, перемножая для каждого из результатов их вероятности и полезности, складывая полученные произведения, а затем вычитая из суммы величину, соответствующую издержкам тестирования. Эта последняя величина высвечивает тот факт, что тесту с низкой валидностью скорее будет отдано предпочтение в ситуации выбора, если он краток, недорог, легко может проводиться малоквалифицированным персоналом и пригоден для группового проведения. Индивидуальному тесту, требующему для своего проведения квалифицированного специалиста или дорогостоящего оборудования, нужно было бы иметь более высокую валидность, чтобы оказаться выбранным для практического использования. В гипотетическом примере на рисунке 13 величина издержек тестирования, оцененных по шкале полезности, составляет 0.10. Общая ожидаемая полезность (*EU*) этой стратегии вычисляется следующим образом:

$$EU = 0.38\times(+1.00) + 0.07\times(-1.00) + 0.33\times0.00 + 0.22\times(-0.50) - 0.10 = +0.10.$$

Полученное значение *EU* можно затем сравнить с другими, вычисленными при различных значениях критического показателя, при применении разных тестов (различающихся по валидности и затратам на проведение и обработку данных) или тестовой батареи, а также при использовании различных стратегий принятия решений.

- **Последовательные стратегии и адаптивный подход**

В некоторых ситуациях эффективность теста можно повысить, применяя более сложные стратегии принятия решений, учитывающие большее число параметров. Во-первых, тесты могут использоваться не только в качестве основания для окончательного решения, но и для последовательного принятия решений. На рисунке 14 показана двухэтапная последовательная стратегия.

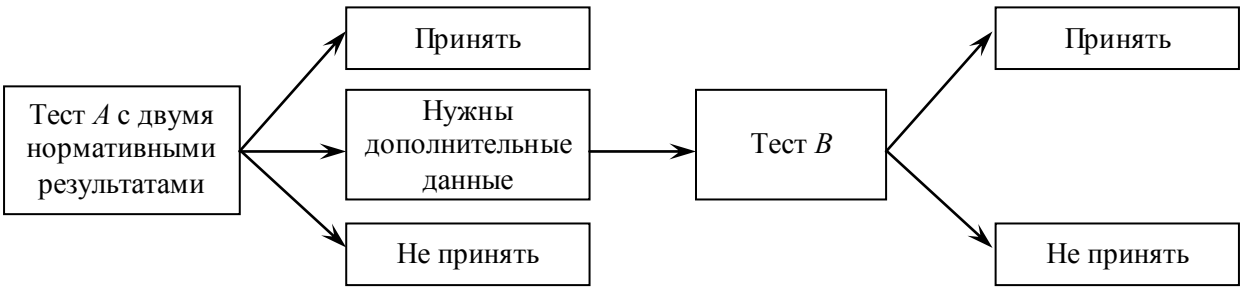


Рисунок 14. Последовательная стратегия принятия решения [2].

В качестве теста *A* можно было бы использовать короткий, легкий в проведении, скрининговый тест. На основе результатов этого теста претендентов можно было бы распределить по трем категориям: те, кто будет принят на работу без дополнительных испытаний; те, кто получит окончательный отказ в приеме, и те, кто образует промежуточную группу «сомнительных» случаев. Далее последних можно было бы подвергнуть более интенсивному обследованию с помощью теста *B*, и уже по результатам второго этапа тестирования разделить эту группу на две категории: принятых и не принятых на работу.

Следует также отметить, что в действительности многие кадровые решения принимаются в соответствии с последовательной стратегией, хотя это и не всегда осознается. Некомпетентные работники, принятые вследствие ошибки прогноза, обычно могут быть уволены по истечении испытательного срока; отчисляются также на ряде этапов не справляющиеся с учебными программами студенты. В таких ситуациях только отрицательное решение оказывается окончательным. Конечно, ошибки отбора, которые затем исправляются, могут дорого обходиться с точки зрения той или иной системы ценностей. Но все-таки они часто сопряжены с меньшими издержками, чем окончательное ошибочное решение.

Вторым условием, влияющим на эффективность психологического теста, является доступность альтернативных методов и возможность адаптивного подхода, учитывающего индивидуальные особенности. Примером может служить использование различных программ и методов подготовки персонала в зависимости от уровня их способностей. В этих условиях стратегия принятия решения в отношении конкретного случая должна строиться с учетом имеющихся сведений о взаимодействии между первоначальным результатом теста и дифференцированным воздействием. Адаптивный подход нередко позволяет значительно повысить процент успешно справляющихся с работой.

Приведенные примеры иллюстрируют лишь несколько областей, в которых понятия и принципы теории принятия решений могут помочь в оценке пригодности психологических тестов для специфических целей тестирования. Знание коэффициента валидности еще недостаточно для ответа на вопрос, следует ли использовать данный тест, поскольку валидность — лишь один из факторов, подлежащих рассмотрению при оценке влияния теста на эффективность всего процесса выработки решений.

- **Объединение данных различных тестов**

Когда несколько специально подобранных тестов применяются вместе для предсказания одного-единственного критерия, такую совокупность тестов называют *тестовой батареей* (*test battery*). Главная проблема, возникающая при использовании таких батарей, касается способа объединения показателей по отдельным тестам при выработке решения в каждом индивидуальном случае. Для этой цели обращаются к двум основным видам процедур, а именно использованию *уравнения множественной регрессии* и *анализу профиля нормативных показателей*. В ряде случаев (например, при оценке кандидатов на ответственные должности) используют обе процедуры для составления заключений или рекомендаций с целью повысить валидность психологических выводов.

Уравнение множественной регрессии позволяет получить числовую оценку прогнозируемого критерия для каждого испытуемого на основе его показателей по всем тестам батареи. Например, при оценке прогнозируемой производственной ценности работника в ходе отбора на ведущие должности энергопредприятия ("Психофизиологические методы", 1979, с. 21) предлагается использовать следующее уравнение множественной линейной регрессии[8]:

$$\text{Производственная ценность} = -13.82 - 0.026\times K + 0.443\times N - 0.023\times t + 0.013\times R + \\ + 0.010\times K_1 + 2.158\times N_1 + 0.038\times t_1 - 0.106\times R_1 - \\ - 0.390\times K_2 + 0.253\times N_2 + 0.039\times t_2 + 0.168\times R_2,$$

где K – коэффициент полезной работы, R – производительность, N – число правильных ответов, t – время выполнения при исследовании оперативного счета; K_1 – коэффициент полезной работы, R_1 – производительность, N_1 – число правильных ответов, t_1 – время выполнения при исследовании технического мышления; K_2 – коэффициент полезной работы, R_2 – производительность, N_2 – число правильных ответов, t_2 – время выполнения при исследовании пространственного мышления.

Предположим, что кандидат получил следующие результаты по тестам: $K = 240$, $R = 206$, $N = 4$, $t = 180$; $K_1 = 260$, $R_1 = 188$, $N_1 = 7$, $t_1 = 270$; $K_2 = 12$, $R_2 = 7$, $N_2 = 34$, $t_2 = 287$. Ожидаемая

[8] Необходимо заметить, что использование множественной линейной регрессии предполагает, что входящие в анализ переменные (критериальный показатель и тестовые данные) нормально распределены (данное ограничение, напомним, распространяется и на использование параметрического коэффициента корреляции Пирсона).

производственная ценность (по пятибалльной шкале) для кандидата определяется тогда суммой произведений значений зависимых переменных на их весовые коэффициенты:

$$\text{Производственная ценность} = -13.82 - 0.026\times240 + 0.443\times4 - 0.023\times270 + 0.013\times206 + 0.010\times260 + 2.158\times7 + 0.038\times270 - 0.106\times188 - 0.390\times12 + 0.253\times34 + 0.039\times287 + 0.168\times7 = 4.58.$$

Уравнение регрессии основано на корреляции каждого теста с критерием и корреляциях тестов между собой. Очевидно, что тесты (независимые переменные или предикторы), сильнее коррелирующие с критерием (зависимой переменной), должны получить больший весовой коэффициент (коэффициент регрессионного уравнения)[9]. Столь же важно, однако, учитывать корреляцию каждого теста с другими тестами батареи. Высокая корреляция указывает на ненужное дублирование одного теста другим, ибо это означает, что тесты в значительной мере направлены на один и тот же аспект критерия. Включение двух таких тестов не повышает существенно валидности всей батареи, даже если оба они тесно коррелируют с критерием. В этом случае один из этих тестов столь же эффективен, как и пара, поэтому в батарее следует оставить только один тест.

Однако даже после того, как случаи наиболее выраженного дублирования тестов в батарее устраняются, оставшиеся тесты все равно будут в той или иной степени коррелировать друг с другом. Для максимизации прогнозирующей силы тесты, вносящие более «уникальный» вклад в полную батарею, должны получать больший вес по сравнению с тестами, частично дублирующими функции других тестов батареи. При расчете коэффициентов уравнения множественной регрессии каждый тест получает вес, прямо пропорциональный его корреляции с критерием и обратно пропорциональный корреляции с другими тестами. Это значит, что максимальный вес получит тест, обладающий наибольшей валидностью и в наименьшей степени дублирующий остальную часть батареи.

Валидность полной батареи можно найти путем вычисления коэффициента множественной корреляции (R) между входящими в нее тестами и критерием. Этот вид корреляции дает оценку максимальной предсказуемостной эффективности, которой можно добиться от данной тестовой батареи при условии, что каждый входящий в нее тест получает

[9] Стандартизированные регрессионные коэффициенты (β), рассчитанные по стандартизированным значениям переменных, позволяют сравнить и оценить значимость независимых переменных, так как β-коэффициент показывает на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной при условии постоянства остальных независимых переменных.

оптимальный — с точки зрения предсказания критерия — вес. Оптимальные веса как раз и определяются по уравнению регрессии.

Коэффициент множественной корреляции уравнения регрессии может быть использован для деления кандидатов на различные группы по результатам батареи тестов: например, оперативник – не оперативник, руководитель – исполнитель и т.д. Для этих целей также можно воспользоваться дискриминантным анализом.

Анализ профиля и нормативных показателей состоит в использовании системы минимальных проходных результатов (нормативов), устанавливаемых для каждого из теста батареи. Когда применяется строгий вариант этого метода, всякий, кто не достигает такого минимального уровня хотя бы по одному из тестов, считается не прошедшим тестирования. При выборе тестов и установлении нормативных показателей, подходящих для определенной профессии, обычно исходят не только из величины коэффициентов валидности тестов. Если бы в расчет принимались только тесты со значимыми коэффициентами валидности, то могли оказаться неучтенными существенные навыки или способности, которыми должны обладать все представители определенной профессии. Поэтому необходимо рассматривать и те способности, которые должны быть хорошо развиты у тестируемых как единой профессиональной группы, даже если индивидуальные различия между ними, наблюдающиеся выше критериального минимума, никак не связаны с успешностью работы. Кроме того, представители некоторых профессий могут представлять собой настолько однородную группу по ключевой переменной, что диапазон индивидуальных различий оказывается слишком узким, чтобы обеспечить значимую корреляцию между показателями теста и критерием.

Наиболее сильный аргумент в пользу применения множественных нормативных показателей, а не уравнения регрессии, основывается на возможности существования компенсирующих показателей (*compensatory scores*). Другими словами, серьезная недостаточность в одном навыке может остаться незамеченной в суммарном показателе индивидуума по тестовой батарее вследствие высокого показателя по другому тесту. Если эта недостаточность относится к навыку, который является решающим для выполнения определенной работы, отобранный кандидат потерпит неудачу, независимо от его способностей в других областях. Однако такой ситуации можно избежать, установив один или несколько критических навыков, необходимых в определенной профессии, и применяя критический показатель только в соответствующих тестах. В большинстве же тестов

обычно предпочтительнее сохранять актуальный, фактический показатель, поскольку, чем выше тестовый показатель конкретного человека, тем выше, в общем, будет эффективность его работы. Для большинства профессий связь между прогнозирующим показателем и критериальной деятельностью носит линейный характер. При этих условиях отбор персонала на основе фактической величины тестовых показателей обеспечивает более высокую эффективность работы, чем отбор на основе превышения минимальных критических показателей.

В методических рекомендациях по отбору в аварийно-спасательные службы [17] приведены следующие нормативные показатели по ряду психологических тестов (см. табл. 19):

Методика, показатель	**Среднее значение**	**Верхняя граница нормы**	**Нижняя граница нормы**	**Недопустимая величина показателя**
Тест Равена				
• Количество правильных ответов	36	53	28	< 25
• Количество ошибок, %	25	0	40	> 40
Простая сенсомоторная реакция				
• Среднее время реакции, мс	236	296	176	> 400
Сложная сенсомоторная реакция				
• Среднее время реакции, мс	366	501	233	> 700
• Количество пропусков	0	1	0	> 2
• Количество ошибочных реакций	1	3	0	> 4

Таблица 19. Нормативы для психодиагностических методик [17].

ДИСКРИМИНАТИВНОСТЬ ТЕСТА

Еще одной особенностью эффективных тестов является дискриминативность. Достижение удовлетворительного распределения показателей является одной из целей разработчика теста. Очевидна низкая ценность психологического теста, по которому все испытуемые показали одинаковые результаты.

Показатели дискриминативности, как указывает Гилфорд (Guilford), связаны по существу с ранжированием испытуемых.

- **Коэффициент Фергюсона**

Для оценивания дискриминативности тестов рекомендован **коэффициент** δ (*дельта*) **Фергюсона** (Ferguson, 1949):

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - [N^2/(n+1)]}$$

Данную формулу можно представить в другом виде:

$$\delta = \frac{(n+1)\cdot(N^2 - \sum f_i^2)}{n \cdot N^2}$$

где f_i - частота встречаемости каждого результата по тесту (следовательно, $\sum f_i = N$); N – общее число испытуемых; n – количество заданий.

Очевидно, что максимум дискриминативности достигается тогда, когда все результаты по тесту имеют одинаковую частоту. В общем виде, коэффициент δ Фергюсона - это отношение между показателем дискриминативности, полученным для некоторого теста, и максимальным значением дискриминативности, которое может обеспечить такой тест. δ = 0, когда все испытуемые получили одинаковые показатели (то есть, когда нет дискриминативности), и δ = 1 при равномерном (прямоугольном) распределении. В таблице 20 приведен пример расчета коэффициента δ Фергюсона.

Номер результата по тесту	**Частота встречаемости результата в выборке (f_i)**	f_i^2
1	2	4
2	4	16
3	5	25
4	8	64
5	15	225
…	…	…
15	3	9
16	1	1
		$\sum f_i^2 = 2742$
$\delta = \frac{(n+1)\cdot(N^2 - \sum f_i^2)}{n \cdot N^2} = \frac{(30+1)\cdot(100^2 - 2742)}{30 \cdot 100^2} = 0.750$		

Таблица 20. Пример расчета коэффициента δ Фергюсона.

(Число заданий n = 30; число обследованных N = 100.)

Разработчик тестов должен учитывать некоторые характеристики коэффициента δ Фергюсона. Поскольку для равномерного (прямоугольного) распределения (наиболее дискриминативного) необходимы задания, в которых бы наиболее полно были реализованы все возможные проявления измеряемого свойства, это означает, что дискриминативность до некоторой степени противостоит надежности (внутренней согласованности тестовых заданий). Распределение значений показателя по тесту это функция трудности заданий и их взаимной коррелированности, а это влияет не только на надежность, но также и на дискриминативность. Из всего этого следует, замечает П. Клайн, что при конструировании теста его назначение определяет то, до какой степени нашей целью является достижение максимальной надежности или максимальной дискриминативности (что важно, например, при профотборе).

- **Коэффициент дискриминации – бисериальный коэффициент корреляции**

Данный подход основывается на анализе способности отдельных пунктов (заданий) теста дифференцировать обследуемых относительно «максимального» или «минимального» результата теста. Пусть ответы испытуемых на конкретное задание теста можно представить в двухбалльной шкале - «верно» (1 балл), «неверно» (0 баллов). Сумма баллов

по всем заданиям представляет собой первичную («сырую») оценку для каждого субъекта. Мера соответствия успешности выполнения одной задачи к оценке по всему тесту является показателем дискриминативности задания теста для данной выборки испытуемых, которая вычисляется в виде *точечно-бисериального коэффициента корреляции* и называется ***коэффициентом дискриминации*** (***индексом дискриминации***)

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_x} \cdot \sqrt{\frac{N_1 \cdot N_0}{(N-1) \cdot N}}$$

где $\bar{x}_1$ - среднее арифметическое оценок по тесту у испытуемых, правильно выполнивших задание (в случае опросника личностного - соответствие с «ключом»); $\bar{x}_0$ - среднее арифметическое оценок по тесту у испытуемых, неверно выполнивших задание; σ_x - среднеквадратическое отклонение индивидуальных оценок по тесту для выборки; N_1 - число испытуемых, правильно решивших задачу (или тех, чей ответ на данный пункт опросника соответствует «ключу»); N_0 - число испытуемых, неверно решивших задачу; N – общее число испытуемых ($N=N_1+N_0$).

Можно привести ряд других эквивалентных выражений расчета бисериального коэффициента корреляции (где $\bar{x}$ - среднее арифметическое всех индивидуальных оценок по тесту):

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}}{\sigma_x} \cdot \sqrt{\frac{N_1 \cdot N}{N_0 (N-1)}}, \qquad r_{pb} = \frac{\bar{x} - \bar{x}_0}{\sigma_x} \cdot \sqrt{\frac{N_0 \cdot N}{N_1 (N-1)}}.$$

В таблице 21 приведен пример вычисления r_{pb} при анализе дискриминативности отдельного пункта личностного опросника, т.е. корреляция между типичным ответом на отдельный пункт (утверждение – отрицание) с общим результатом по тесту. Вычисленное таким образом значение r_{pb}(=0.46) показывает, что проверяемый пункт опросника имеет среднюю диагностическую значимость и слабо коррелирует с общим результатом по тесту.

Коэффициент дискриминации может принимать значение от –1 до +1. Высокий положительный r_{pb} свидетельствует об эффективности деления испытуемых. Высокое отрицательное значение r_{pb} свидетельствует о непригодности данной задачи для теста, о ее несоответствии суммарному результату.

Коэффициент дискриминативности заданий теста является, по сути, показателем критериальной валидности отдельного пункта, поскольку определяется по отношению к внешнему критерию - суммарному результату или оценкам продуктивности реальной

деятельности испытуемых.

Номер обследуемого	Ответ на пункт опросника (Y)	Результат по шкале в "сырых" баллах (X)	Вычисление
1	1	16	$N_1 = 11;\ N_0 = 7;\ N = 18$
2	0	12	
3	0	11	$\bar{x}_1 = 12.36;\quad \bar{x}_0 = 10.00$
4	1	7	
5	1	15	$\sigma_x = 2.55$
6	1	14	$\sigma_{\tilde{o}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$
7	0	10	
8	0	11	
9	1	15	
10	0	9	$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_x} \cdot \sqrt{\frac{N_1 \cdot N_0}{(N-1) \cdot N}} =$
11	1	13	
12	0	7	$\frac{12.36 - 10.00}{2.55} \cdot \sqrt{\frac{11 \cdot 7}{(18-1) \cdot 18}}$
13	1	13	
14	1	11	
15	0	10	$r_{pb} = 0.46$
16	1	11	
17	1	10	
18	1	11	
Примечание: 1 – совпадение с "ключом"; 0 – несовпадение с "ключом".			

Таблица 21. Пример вычисления индекса дискриминации отдельного пункта личностного опросника.

- **Индекс дискриминации (*D*)**

Данный индекс вычисляется с применением метода контрастных групп. Необходимым условием применения метода в этом случае является *наличие близкого к нормальному распределения оценок по критерию валидизации*.

Доля членов контрастных групп может изменяться в широких пределах в зависимости от величины выборки. Чем больше выборка, тем меньшей долей испытуемых можно ограничиться при выделении групп с высоким (N_{max}) и низким результатами (N_{min}). Нижняя

граница «отсечения групп» составляет 10% общего числа испытуемых в выборке, верхняя - 33%. 10% группы берутся редко, поскольку их малочисленность снижает статистическую надежность индексов дискриминации. Чаще из выборки «извлекают» по 27 или 33% испытуемых.

Индекс дискриминации вычисляется как разность между долей лиц, правильно решивших задачу из «высокопродуктивной» ($N_{n_{\max}}$) и «низкопродуктивной» групп ($N_{n_{\min}}$), и обозначается ***D***:

$$D = \frac{N_{n_{\max}}}{N_{\max}} - \frac{N_{n_{\min}}}{N_{\min}}$$

Поскольку $N_{max} = N_{min}$ ($0.10N \div 0.33N$), то уравнение приобретает вид:

$$D = \frac{N_{n_{\max}} - N_{n_{\min}}}{(0.10 \div 0.30) \cdot N}$$

Индекс дискриминации изменяется в пределах от –1 до +1. Для оценки заданий по величине индекса дискриминации можно предложить следующую таблицу (см. табл. 22):

Величина индекса дискриминации (*D*)	**Интерпретация**
≥ 0.40	Задание вполне эффективное.
0.30 ÷ 0.39	Задание удовлетворительное.
0.20 ÷ 0.29	Задание следует проанализировать на пригодность использования в тесте.
< 0.19	Задание необходимо изъять или тщательно проанализировать и переработать. Оно практически не обладает дифференцирующей способностью.
< 0.00	Задание некачественное, так как лучшая группа студентов отвечает на него хуже, чем слабая.

Таблица 22. Общие принципы для интерпретации индекса дискриминации (*D*).

- **Четырехпольный коэффициент корреляции**

Для расчета дискриминативности заданий теста также можно воспользоваться

четырехпольным коэффициентом корреляции (четырехпольный коэффициент ассоциации Пирсона φ для переменных, измеряемых по дихотомической шкале наименований):

$$r_{phi} = \frac{f_g - f_d}{\sqrt{pq}}$$

где f_g - число лиц, правильно решивших задачу, по отношению к общему числу обследованных в группе с максимальным результатом; f_d – число лиц, правильно решивших задание, по отношению к общему числу обследованных в группе с минимальным результатом; p - общая пропорция (f_g + f_d) правильно выполнивших задание; q - число лиц, давших неверное решение (1 - p).

Критические значения этого коэффициента, свидетельствующие о диагностической ценности (на уровне p < 0.05) в зависимости от количества обследованных (N), приведены ниже:

N	25	50	100	200
r_{phi}	0.39	0.28	0.20	0.14

Максимальная точность определения r_{phi} достигается в случае, когда максимальная и минимальная группы составляют по 27% выборки.

При анализе дискриминативности заданий теста особое внимание следует уделить определению статистической значимости коэффициентов корреляции. В тех случаях, когда значение коэффициента дискриминации приближается к нулю и уровень значимости невысок, проверяемый пункт теста должен быть пересмотрен в связи с некорректностью формулировки задания или вариантов ответов на него.

Определение дискриминативности обязательно для тестов, влияющих на итоговую аттестацию, оценку специалиста, на отбор кандидатов и распределение по должностям.

- **Коэффициент корреляции Гилфорда – дискриминативность теста**

Фи - коэффициент корреляции Гилфорда, рассмотренный нами ранее для оценки надежности параллельных форм критериально-ориентированного теста (см. Коэффициент корреляции Гилфорда и каппа-коэффициент), может быть использован и для определения дискриминативности теста методом контрастных групп. Допустим, по оценкам экспертов

мы сформировали две группы обучаемых: с высокой тревожностью в процессе выполнения ответственных заданий и с низкой тревожностью. Результаты теста на тревожность также позволили нам выделить две группы с высокими и низкими показателями тревожности (см. табл. 23):

	Оценки экспертов	
Результаты теста	**Высокая тревожность**	**Низкая тревожность**
Высокая тревожность	*a*	*b*
Низкая тревожность	*c*	*d*

Таблица 23. Оценки экспертов и результаты теста на тревожность.

На основе матрицы сопряженности 2 × 2 вычисляется φ - коэффициент корреляции Гилфорда для оценки дискриминативности теста:

$$\varphi = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

- **Коэффициент корреляции Гилфорда – дискриминативность заданий**

Рассмотрим использование φ - коэффициента корреляции Гилфорда для оценки дискриминативности отдельных заданий теста с дихотомической шкалой ответа "Верно-Неверно". В этом случае суммарный балл по тесту позволяет выделить две экстремальные группы с "высокими" и "низкими" показателями (10÷33% от общего числа испытуемых в каждой группе) и построить следующую матрицу сопряженности 2 × 2 для каждого задания (см. табл. 24).

	Экстремальная группа	
Ответ	**Высокая**	**Низкая**
"Верно"	*a*	*b*
"Неверно"	*c*	*d*

Таблица 24. Матрица сопряженности ответов на задание по каждой группе.

Очевидно, что "хорошее" задание (вопрос, пункт опросника), обладающее высокой дискриминативностью относительно экстремальных групп, должно обладать высоким контрастом значений *a* и *b*, с одной стороны, и одновременно высоким контрастом *c* и *d* — с другой. При этом контрасты должны иметь противоположный знак: если *a* больше *b* («верно» чаще отвечает «высокая» группа), то *c* должно быть меньше *d* («неверно» чаще отвечает «низкая» группа). Если мы анализируем пункты опросника, то в случае «обратных» пунктов (вопросов), когда разность *a*–*b* отрицательна, то разность *c*–*d* должна быть положительна.

Для анализа представленной матрицы сопряженности предлагают использовать следующий вариант φ - коэффициента корреляции Гилфорда [19]:

$$\varphi = \frac{ad - bc - 0.5 \cdot N}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \quad (1)$$

где *N* – сумма всех элементов таблицы: $N = a + b + c + d$.

Для оценки дискриминативности «обратных» пунктов (вопросов) данная формула имеет следующий вид:

$$\varphi = \frac{ad - bc + 0.5 \cdot N}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \quad (2)$$

Использование равных и известных по объему экстремальных групп $(a + c) = (b + d)$ позволяет несколько упростить представленные формулы:

$$\varphi = \frac{a \cdot N - P \cdot S - 0.5 \cdot N}{\sqrt{P \cdot (N - P)}} \quad (1a)$$

$$\varphi = \frac{a \cdot N - P \cdot S + 0.5 \cdot N}{\sqrt{P \cdot (N - P)}} \quad (2a)$$

где $P = a + b$; $S = (a + c) = (b + d)$.

Для оценки значимости φ - коэффициента используем ранее рассмотренное соотношение:

$$\left|\varphi_{\hat{e}\eth\grave{e}\grave{o}}\right| = \sqrt{\frac{\chi^2}{N}}$$

Пусть $N = 100$, тогда для уровня значимости $p = 0.01$ табличное значение $\chi^2 = 6.62$ и $\varphi_{крит} = 0.257$. Если вычисленное значение φ - коэффициента для конкретного пункта опросника

превышает по модулю $\varphi_{крит}$, то с вероятностью ошибки в 1% можно утверждать, что данный пункт вносит значимый вклад в суммарный балл теста. Если $\varphi > 0.257$, пункт следует считать «прямым» (ответ «верно» свидетельствует в пользу измеряемой черты), если $\varphi < -0.257$, пункт следует считать «обратным» (ответ «неверно» свидетельствует в пользу измеряемой черты).

После расчета φ - коэффициентов корреляции Гилфорда все задания (вопросы, пункты опросника), для которых были получены незначимые значения коэффициентов корреляции, удаляются из теста (или переформулируются) и процедура повторяется, пока в тесте не останутся только те задания, которые имеют значимые коэффициенты корреляции Гилфорда.

- **Алгоритм раздельного коррелирования ответов**

А.Г. Шмелев и В.И. Похилько [19] предложили свой алгоритм оценки дискриминативности заданий (пунктов опросников), который представляет собой развитие представленного выше подхода Гилфорда. Алгоритм раздельного коррелирования ответов позволяет учитывать при подсчете суммарного балла по тесту разные веса (ключи) ответов «верно» и «неверно», их разное диагностическое значение.

Как и при расчете φ - коэффициента корреляции Гилфорда, на первом шаге по суммарному показателю теста выделяются «высокая» и «низкая» группы, и для каждого пункта j опросника строится матрица сопряженности 2 × 2. Ключ для ответа «верно» определяется по формуле:

$$f_j^+ = \frac{a-b}{a+b}$$

где f_j^+ достигает +1, если «верно» отвечают только представители «высокой» группы, и −1, если «верно» отвечают только испытуемые из «низкой» группы. Значимость f_j^+ можно оценить с помощью следующего приближенного соотношения:

$$f_{êðèò} = \sqrt{\frac{\chi^2}{a+b}}$$

Ключ для ответа «неверно» определяется по формуле:

$$f_j^- = \frac{c-d}{c+d}$$

Соответственно значимость f_j^- можно оценить с помощью следующего приближенного соотношения:

$$f_{êðèò} = \sqrt{\frac{\chi^2}{c+d}}$$

Если $f_j^+ > f_{крит}$, то положительному ответу на вопрос ("верно") присваивается ключ, равный f_j^+. Если $f_j^- < f_{крит}$, то отрицательному ответу на вопрос ("неверно") присваивается ключ, равный f_j^-. В противном случае пункт опросника исключается из теста (низкая дискриминативность вопроса).

Практическое испытание данного алгоритма показало, что более высокими f_j, как правило, обладают менее социально одобряемые ответы. Например, в опроснике на «склонность к риску» ответ «верно» на вопрос «Я быстро меняю свои интересы и увлечения» получил f_j^+ = +0.60 (p < 0.05), а ответ «неверно» получил f_j^- = –0.26 (не значимо). Понятно, что устойчивость интересов — более социально одобряемая форма поведения, поэтому более информативен ответ «верно» (контраст между *a* и *b* сильнее контраста между *c* и *d*). По этим же причинам для вопроса «Я быстрей испытываю скуку, чем большинство людей, делающих то же самое» были получены следующие данные: f_j^+ = +0.63, а f_j^- = –0.29. Точно также информативнее оказались менее «благоразумные» ответы. Например, для пункта «Люди слишком часто безрассудно тратят собственное здоровье, переоценивая его запасы» f_j^+ = –0.14, а f_j^- = +0.75 (значимо на уровне p < 0.05). В опроснике на «тревожность» в вопросе «Я опасаюсь, что о моих недостатках станет известно другим» более информативным оказался ответ «неверно»: f_j^+ = +0.33, а f_j^- = –0.78.

Результаты исследований позволили А.Г. Шмелеву и В.И. Похилько прийти к мнению, что алгоритм раздельного коррелирования ответов является более эффективным при отборе пунктов теста-опросника по внешнему критерию, чем расчет φ - коэффициента корреляции Гилфорда.

ЛИТЕРАТУРА

[1] Армстронг М. Практика управления человеческими ресурсами. 8-е издание. СПб.: Питер, 2004, 832 с.

[2] Анастази А. Психологическое тестирование. Книга 1. М., Педагогика, 1982. 320 с.

[3] Анастази А., Урбина С. Психологическое тестирование. СПб.: Питер, 2005. 688 с.

[4] Большой психологический словарь. Под ред. Б. Г. Мещерякова, В. П. Зинченко Изд.: прайм-ЕВРОЗНАК, 2005, 672 стр.

[5] Десслер Г. Управление персоналом. М.: Издательство БИНОМ, 1997. 432 с.

[6] Бурлачук Л.Ф., Морозов С.М. Словарь-справочник по психологической диагностике. Киев: Наук. думка, 1989, 200 с.

[7] Клайн П. Справочное руководство по конструированию тестов: Введение в психометрическое проектирование. Перевод с английского / Под ред. Л.Ф. Бурлачука. Киев: ПАН Лтд., 1994. 288 с.

[8] Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. - М.: ФИЗМАТЛИТ, 2006. - 816 с.

[9] Кондаков И.М. Создание психологических опросников с помощью статистического пакета SPSS for Windows 11.5.0. Учебно-методическое пособие. (http://www.matlab.mgppu.ru/work/0028.htm)

[10] Кондаков И.М., Романюк Э.И., Сорокина О.Л., Шишлянникова Л.М. Разработка тестовых заданий для анализа знаний студентов. Методическое пособие. М.: МГППУ, 2005, 66 с.

[11] Корсини Р., Ауэрбах А. (Ред.) Психологическая энциклопедия. 2-е издание. Питер, 2006, 1096 с.

[12] Одегов Ю.Г., Журавлев П.В. Управление персоналом: Учебник для вузов. М.: Финстатинформ, 1997, 878 с.

[13] Павлов А.Н., Соколов Б.В. Методы обработки экспертной информации. Учебно-методическое пособие. Санкт-Петербург, 2005, 34 с.

[14] Полякова О.Н. Оценка деятельности работников. Материалы к лекциям по курсу "Управление персоналом" и спецкурсу "Оценка деятельности работников". Воронеж, 2001, 38 с.

[15] Психологическая диагностика: Учебное пособие. Под ред. К.М. Гуревича и Е.М. Борисовой. М.: Изд-во УРАО, 1997, 304 с.

[16] Психофизиологические методы профессионального отбора в ведущие профессии энергопредприятий. Методические рекомендации. Киев, 1979, 24 с.

[17] Психофизиологический профессиональный отбор и периодический психофизиологический контроль персонала аварийно-спасательных формирований. Методические рекомендации, М., 1995, 32 с.

[18] Шмелев А.Г. Психодиагностика личностных черт. С.-Петербург: Речь, 2002, 480 с.

[19] Шмелев А.Г., Похилько В.И. Анализ пунктов при конструировании и применении тест-опросников: ручные и компьютерные алгоритмы. Вопросы психологии, 1985, № 4, с. 126-134.

[20] Measurement Validity Types. In The Manual: William M.K. Trochim. Cornell University. The Research Methods Knowledge Base. (www.socialresearchmethods.net/kb/measval.php)

[21] Testing And Assessment: An Employer's Guide To Good Practices. U.S. Department of Labor Employment and Training Administration. 1999. 80 p.

[22] Милкович Д.Т., Ньюман Д. М. Система вознаграждения и методы стимулирования персонала. 2005, 760 с.

[23] Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. Киев, МОРИОН. 2002, 640 с.

ПРИЛОЖЕНИЯ

1. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА

Коэффициент корреляции произведения моментов Пирсона учитывает не только положение индивидуума в группе, но и степень его отклонения в ту или иную сторону от группового среднего значения. Когда положение каждого индивидуума выражается в единицах стандартного показателя ($z = \frac{X - M}{\sigma}$, где M – среднее значение по группе, X – индивидуальный результат, σ - дисперсия групповых данных), те, кто занимает положение выше среднего, получают положительные стандартные показатели, а те, кто находится ниже среднего уровня, - отрицательные. Таким образом, испытуемый, превосходящий группу по уровню обеих коррелируемых переменных, будет иметь два положительных стандартных показателя, а испытуемый, отстающий от группы по уровню этих переменных, - два отрицательных. Если теперь перемножить стандартные показатели каждого из этих испытуемых по обеим переменным, то оба произведения будут положительны. Коэффициент корреляции Пирсона есть просто среднее арифметическое всех таких произведений. Его числовое значение бывает высоким и положительным, когда соответствующие показатели z имеют по обеим переменным одинаковые знаки и приблизительно равную величину. Когда испытуемых занимают положение выше среднего по одной переменной, но ниже среднего по другой, то соответствующие произведения будут отрицательны. А если сумма произведений отрицательна, то отрицательной будет и корреляция. Когда же одни произведения отрицательны, а другие положительны, корреляция будет близка к нулю.

На практике нет необходимости переводить каждый первичный показатель в стандартный перед нахождением их произведений, так как это преобразование можно выполнить разом для всех показателей после суммирования их попарных произведений. Существует много ускоренных методов вычисления коэффициента корреляции Пирсона. В таблице 25 показан один из методов вычисления коэффициента корреляции Пирсона (r) между показателями по "Тесту 1" и "Тесту 2" у 10 обследуемых. В двух столбцах справа от имен учеников приведены их показатели по первому (X) и второму (Y) тесту. Суммы и средние арифметические 10 показателей приведены под соответствующими столбцами. В третьем

столбце приведены отклонения (x) каждого показателя "Теста 1" от среднего арифметического этих показателей, а в четвертом - отклонения (y) индивидуальных показателей по "Тесту 2" от их среднего арифметического. Квадраты этих отклонений даны в следующих двух столбцах таблицы, а суммы квадратов отклонений используются при вычислении стандартных отклонений показателей по обоим тестам. Вместо того чтобы каждое x и y делить на соответствующее σ для получения стандартных показателей, это деление выполняется только раз, в конце, как показано в формуле коэффициента корреляции в нижней части таблицы 25. Попарные произведения (xy) в последнем столбце получены перемножением соответствующих отклонений в столбцах (x) и (y). Для вычисления корреляции (r) сумма этих попарных произведений делится на число случаев (N) и на произведение двух стандартных отклонений ($\sigma_x \cdot \sigma_y$).

Обследуемые	**Тест 1 (*X*)**	**Тест 2 (*Y*)**	***X*–*M* (*x*)**	***Y*–*M* (*y*)**	x^2	y^2	***xy***
1	41	17	+ 1	– 4	1	16	– 4
2	38	28	– 2	+ 7	4	49	– 14
3	48	22	+ 8	+ 1	64	1	8
4	32	16	– 8	– 5	64	25	40
5	34	18	– 6	– 3	36	9	18
6	36	15	– 4	– 6	16	36	24
7	41	24	+ 1	+ 3	1	9	3
8	43	20	+ 3	– 1	9	1	– 3
9	47	23	+ 7	+ 2	49	4	14
10	40	27	0	+ 6	0	36	0
$\sum =$	**400**	**210**	**0**	**0**	**244**	**186**	**86**
M =	**40**	**21**					

$$\sigma_x = \sqrt{\frac{x^2}{N}} = \sqrt{\frac{244}{10}} = 4.94; \quad \sigma_y = \sqrt{\frac{y^2}{N}} = \sqrt{\frac{186}{10}} = 4.31;$$

$$r_{xy} = \frac{\sum_{i=1}^{N}[(x_i - \bar{x})(y_i - \bar{y})]}{N \cdot \sigma_x \cdot \sigma_y} = \frac{\sum xy}{N \cdot \sigma_x \cdot \sigma_y} = \frac{86}{10 \cdot 4.94 \cdot 4.31} = 0.40.$$

Таблица 25. Вычисления коэффициента корреляции Пирсона.

Для автоматизированного расчета коэффициента корреляции Пирсона более удобна следующая формула:

$$r_{xy} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{[N\sum x_i^2 - (\sum x_i)^2]\cdot[N\sum y_i^2 - (\sum y_i)^2]}$$

Коэффициент корреляции Пирсона представляет собой меру линейной зависимости двух переменных. Если возвести его в квадрат, то полученное значение коэффициента детерминации (r^2_{xy}) представляет долю вариации, общую для двух переменных ("степень" зависимости или связанности двух переменных).

Используя коэффициент корреляции Пирсона следует учитывать, что он основан на предположении о нормальности распределения исследуемых переменных (параметрический критерий). Если распределение переменных отличается от нормального, то для коэффициента корреляции Пирсона нельзя применять методы проверки на значимость. Другим фактором, часто ограничивающим применимость данного критерия, является объем или размер выборки, доступной для анализа. Поскольку на малой выборке нет способа проверить предположение, что выборочное распределение нормально, мы не можем быть уверенны в возможности использования коэффициента корреляции Пирсона. Также коэффициент корреляции Пирсона не очень устойчив к выбросам. При их наличии можно ошибочно сделать вывод о существовании корреляции между переменными. Поэтому, если распределение исследуемых переменных отличается от нормального[10] или возможны выбросы, то лучше воспользоваться его непараметрическим (свободным от формы распределения переменных) аналогом - коэффициентом ранговой корреляции Спирмена, либо коэффициентом ранговой корреляции Кендалла.

Последнее замечание относится к проблеме измерения. Например, эксперт оценивает компетенцию руководителя "Принятие решений" по 7-ми бальной шкале. Можно ли при этом утверждать, что различие между компетенцией руководителя *A* и *B* (6 и 4 балла) сравнимы с различием между руководителями *D* и *C* (5 и 3 балла)? Анализ оценок экспертов показывает, что они скорее упорядочивают руководителей по уровню выраженности ("низкая" – "высокая") оцениваемой компетенции. Эти измерения

[10] Для проверки нормальности распределения переменной используются критерии χ^2 и Колмогорова-Смирнова.

выполнены в порядковой шкале, а не интервальной, в которой интервалы можно разумным образом сравнивать между собой (например, $A - B = D - C$). Коэффициент корреляции Пирсона предполагает, что исходные данные заданы в интервальной шкале. Для шкал порядка могут быть использованы ранговые коэффициенты корреляции. В тех случаях, когда мы сомневаемся в точности измерения, в том, что мы имеем дело с данными в интервальной шкале, необходимо использовать ранговые коэффициенты корреляции.

2. КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

При расчете непараметрического коэффициента ранговой корреляции Спирмена (r_s) значения переменных должны быть представлены в порядковой шкале (например, рангами). Коэффициент корреляции Спирмена может использоваться для оценки зависимости между переменными независимо от их распределения. Также он менее чувствителен к выбросам, что является ещё одним важным качеством при обработке экспериментальных данных.

Коэффициент корреляции рангов Спирмена (r_s) определяется из уравнения:

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n},$$

где d_i - разности между рангами каждой переменной из пар значений X и Y; n - число сопоставляемых пар.

Если ранжировки экспертов имеют связанные ранги, то коэффициент корреляции Спирмена вычисляется по формуле:

$$r_s^{'} = \frac{r_s - T_x - T_y}{\sqrt{(1 - 2 \cdot T_x) \cdot (1 - 2 \cdot T_y)}},$$

где T_x и T_y (показатели связанных рангов по каждому эксперту) рассчитываются по формулам:

$$T_x = \frac{1}{2 \cdot (n^3 - n)} \cdot \sum_{k=1}^{H_x} (h_k^3 - h_k), \; T_y = \frac{1}{2 \cdot (n^3 - n)} \cdot \sum_{k=1}^{Hy} (h_k^3 - h_k),$$

где h_k – объем связанных групп по каждому эксперту.

В таблице 26 представлен пример вычисления коэффициента ранговой корреляции Спирмена.

Обследуемые	**Тест 1 (*X*)**	**Тест 2 (*Y*)**	**Ранг *X***	**Ранг *Y***	d_i	d^2_i	S_1	S_2
1	1.20	15	1	**4**	– 3	9	5	2
2	1.00	15	**2.5**	**4**	– 1.5	2.25	5	2
3	1.00	18	**2.5**	1	1.5	2.25	7	0
4	0.91	15	**4.5**	**4**	0.5	0.25	5	1
5	0.91	13	**4.5**	**9**	– 4.5	20.25	0	3
6	0.90	13	6	**9**	– 3	9	0	3
7	0.88	17	7	2	5	25	3	0
8	0.86	14	8	**6.5**	1.5	2.25	1	0
9	0.76	14	9	**6.5**	2.5	6.25	1	0
10	0.75	13	10	**9**	1	1	0	0
H_j			2	3	$\sum =$	77.5	**27**	**11**
h_k			$h_1 = 2$ $h_2 = 2$	$h_1 = 3$ $h_2 = 3$ $h_3 = 2$	r_s - несвязанные ранги: $r_s = 1 - \frac{6 \cdot 77.5}{10 \cdot (100-1)} = 0.530$			
$\sum_{k=1}^{H_j}(h_k^3 - h_k)$			6+6 =12	24+24+6 = 54	r_s - связанные ранги: $r_s' = \frac{0.530 - 0.006 - 0.027}{\sqrt{(1-2\cdot 0.006)\cdot(1-2\cdot 0.027)}} = 0.514$			
T_i			0.006	0.027				

Таблица 26. Вычисления коэффициента ранговой корреляции Спирмена.
(Цветным шрифтом выделены связанные ранги для каждой эксперта.)

3. КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ КЕНДАЛЛА

Другим непараметрическим критерием оценки зависимости между переменными служит коэффициент корреляции рангов Кендалла τ (*may*), который определяется следующей формулой:

$$\tau = \frac{S}{0.5 \cdot (n^2 - n)},$$

где $S = S_1 - S_2$; n - число сопоставляемых пар.

Для подсчета S_1 и S_2 (см. табл. 26) данные по "Тесту 1" (X) отсортировываются в возрастающем порядке. Далее анализируются данные (ранги) по "Тесту 2" (Y).

- Для каждого обследуемого подсчитывается, сколько его ранг по Y меньше, чем ранги обследуемых, расположенных ниже (колонка S_1). Полученные значения суммируются по всем обследуемым. Чем выше значение показателя S_1, тем выше степень совпадения последовательности рангов Y с последовательностью рангов X.
- Для каждого обследуемого подсчитывается, сколько его ранг по Y больше, чем ранги обследуемых, расположенных ниже (колонка S_2). Полученные значения суммируются по всем обследуемым. Чем выше значение показателя S_2, тем ниже степень совпадения последовательности рангов Y с последовательностью рангов X.

Подставляя полученные данные в формулу ($S_1 = 27$, $S_2 = 11$), получаем:

$$\tau = \frac{27-11}{0.5\cdot(100-10)} = 0.356 .$$

Если ранжировки экспертов имеют связанные ранги, то коэффициент корреляции Кендалла вычисляется по формуле:

$$\tau' = \frac{S}{\sqrt{0.5\cdot(n^2-n)-T_x}\cdot\sqrt{0.5\cdot(n^2-n)-T_y}},$$

где T_x и T_y (показатели связанных рангов по каждому эксперту) рассчитываются по формулам:

$$T_x = \frac{1}{2}\cdot\sum_{k=1}^{H_x}(h_k^2-h_k),\ \ T_y = \frac{1}{2}\cdot\sum_{k=1}^{Hy}(h_k^2-h_k).$$

Согласно таблице 26:

$$T_x = 0.5*(2+2) = 2;\ T_y = 0.5*(6+6+2) = 7.$$

$$\tau' = \frac{27-11}{\sqrt{0.5\cdot(100-10)-2}\cdot\sqrt{0.5\cdot(100-10)-7}} = 0.396$$

Заметим, что статистики Кендалла (τ) и Спирмена (r_s) имеют различную интерпретацию. Если статистика Спирмена может рассматриваться как прямой аналог статистики Пирсона (r), вычисленный по рангам, статистика Кендалла скорее основана на вероятности. Более точно, проверяется, что имеется различие между вероятностью того, что наблюдаемые данные расположены в том же самом порядке для двух величин и вероятностью того, что они расположены в другом порядке.

4. ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

После расчета коэффициента корреляции между двумя переменными необходимо оценить его значимость: отличается ли полученное значение от нуля и не является ли оно следствием только выборочной ошибки. Когда корреляция для данных выборки экспертов значима на уровне $p = 0.01$ (или 1%), то это означает, что существует не более одного шанса из ста, что в общей популяции экспертов данный коэффициент равен нулю. Уровни значимости указывают риск ошибки, на который мы вынуждены пойти, делая выводы из полученных данных. Если корреляция значима на уровне $p = 0.05$, то вероятность ошибки составляет 5 из 100. В большинстве психологических исследований применяются уровни значимости 0.01 и 0.05.

5. АНАЛИЗ РАБОТЫ (СОДЕРЖАТЕЛЬНАЯ ВАЛИДНОСТЬ)

Анализ содержания работы (*job analysis*) - это исследовательский процесс, определяющий наиболее существенные составные части работы. Его цель заключается в выявлении существенных характеристик работы и требований к исполнителям, необходимых для выполнения данной работы. Таким образом, анализ работы имеет два аспекта:

1. Анализ с ориентацией на задачу - для определения обязанностей, ответственности, методов выполнения работы и т.д. Эти данные используются при создании ***описания работы*** (*job description*) – в чем заключается работа.

2. Анализ с ориентацией на работника - для определения характеристик поведения работника, требуемых для успешного выполнения им своих обязанностей. Эти данные используются при формулировании ***спецификаций работы*** (*job specifications*) - каких людей на нее нанимать.

Чтобы быть эффективным, анализ содержания работы должен установить требования, которые отличают определенный вид работы от всех других. Для получения достаточно полной картины конкретной профессиональной деятельности, аналитик может воспользоваться опубликованными руководствами по обучению конкретной профессии или должностными инструкциями, официальными отчетами о выполнении определенных видов работ и, что особенно важно, может получить консультацию экспертов в данной области - инструкторов производственного обучения, опытных работников и их непосредственных руководителей[11]. При этом в состав экспертной группы не должны входить работники, не имеющие релевантного опыта работы.

Анализ начинается с составления полного перечня работ. Затем тщательно изучается их содержание путем расчленения работы на отдельные элементы, решаемые задачи, логические ступени их выполнения, круг обязанностей. Изучаются приемы и методы, материалы, инструменты и оборудование, с помощью которых выполняется работа; выявляются условия труда, в которых она совершается. После этого устанавливаются объем знаний, мастерство и способности, необходимые для выполнения работы на требуемом уровне.

[11] Эффективным инструментом для получения экспертных заключений может служить техника номинальной группы.

Ниже перечислены основные информационные блоки, на основе которых разрабатываются, например, вопросы для проведения интервью, составляются анкеты, нацеленные на анализ работы [12]:

- Место работы (название компании, подразделение).
- Название должности.
- Должность руководителя работника.
- Должности персонала, находящегося в непосредственном подчинении у работника.
- Главные цели работы.
- Перечень основных задач (их регулярность и относительная важность) и обязанностей работника, размер ответственности.
- Основные показатели работы и стандарты оценки: торговый оборот, размер контролируемых финансовых ресурсов, количество персонала, производительность и т.п.
- Сведения об использовании специального оборудования или станков, приемы работы.
- Информация о необходимости работы с людьми (в или за пределами организации).
- Условия труда: физические условия; рабочие часы, дни и отпуска; социальные условия; экономические условия.
- Особые обстоятельства, такие, как неудобные часы работы, командировки, неблагоприятные и опасные условия труда.
- Требуемое образование и профессиональная квалификация: минимальный и желательный уровень.
- Необходимость обучения, требуемый опыт (минимальный и желательный).
- Возможности: продвижения, перемещения, обучения (развития) навыкам, расширения содержания работы.
- Особые навыки или способности (физические и психологические характеристики, требуемые от индивидуума при выполнении работы): например, умение работать с цифрами, способность четко говорить и т.д.

Согласно М. Армстронгу [1], анализ работ дает следующие данные о деятельности:

- *Общая цель* — для чего существует данная должность и, в сущности, какой вклад ожидается от занимающего ее работника.

- *Содержание* (характер и сфера работы) — выполняемые задачи, операции и обязанности, то есть процессы преобразования инвестиций (знания, навыки и способности) в продукты (результаты).
- *Подотчетность* — результаты или продукты, за которые работник на этой должности отвечает.
- *Критерии выполнения* — критерии, меры или показатели, которые дают возможность оценить выполненную работу.
- *Ответственность* — ответственность работника, занимающего должность, относительно масштаба и вложений в работу.
- *Организационные факторы* — подчинение работника, занимающего данную должность, то есть кому он/она подчиняется непосредственно (линейный менеджер) или функционально (по вопросам, связанным со специализированными областями, такими, как управление финансами или персоналом); кто прямо или косвенно подчиняется данному работнику; в какой степени он включен в групповую работу.
- *Принятие решений* — объем полномочий при принятии решений; сложность, масштаб, разнообразие и запутанность решаемых проблем.
- *Ресурсы* — количество и стоимость ресурсов, находящихся под его управлением.
- *Коммуникация* — тип и важность межличностных отношений.
- *Мотивационные факторы* — конкретные особенности работы, которые могут мотивировать или демотивировать работника.
- *Факторы развития* — повышение в должности и карьерные перспективы, возможность приобрести новые навыки или специальные знания.
- *Факторы среды* — условия труда, охрана труда и техника безопасности, работа в ночное или вечернее время, подвижность и эргономические факторы, связанные с дизайном и использованием оборудования или рабочих мест.

Гари Десслер приводит следующие типы информации, которые собираются в процессе анализа содержания работы [5]:

- *Рабочая деятельность.* Прежде всего, обычно собирается информация о реальных видах рабочей деятельности, таких, как чистка, шитье или рисование. Иногда такой список содержит описание того, как, почему и когда работник выполняет каждый из видов работ.

- *Человеческое поведение.* Может быть включена информация о человеческом поведении, таком, как чувственность, общение, принятие решений и творческие навыки. Сюда должна быть включена информация относительно требований, предъявляемых непосредственно к

человеку, в терминах затрат человеческой энергии, необходимости ходить на длинные дистанции и т. д.

- *Механизмы, оборудование, инструменты и другие приспособления, используемые в работе.* В этот раздел следует включить данные, касающиеся производимых продуктов, обрабатываемых материалов, знаний, которые должны применяться (такие, как знание законов физики), а также оказываемых услуг (таких, как консультирование или ремонт).

- *Нормы производительности.* Собирается информация о нормах производительности (в терминах количества, качества или затрачиваемого времени на каждый вид работы), нормах и критериях, по которым будет оцениваться работа.

- *Рабочее окружение.* В этот раздел следует включить информацию, касающуюся таких аспектов, как физические условия работы, расписание работы, а также организационное и социальное окружение — например, люди, с которыми работнику придется общаться в процессе работы. Также в этот раздел может быть включена информация о финансовых и нефинансовых стимулах.

- *Требования к человеку.* Обычно приводится информация, касающаяся таких требований к потенциальному работнику, как знания и умения (образование, обучение, опыт работы и т. д.) и требуемые личные характеристики (интересы, склонности, способности, физические данные и т. п.).

В американских справочниках профессий описание каждой работы содержит не только определение обязанностей, ответственности, методов выполнения работы и т.п. (как в российском «Квалификационном справочнике должностей руководителей, специалистов и других служащих»), но и определение характеристик поведения работника, требуемых для успешного выполнения им своих обязанностей (примером может служить Национальный Профессиональный квалификационный справочник США - NVQ). Анализ требований к исполнителю (*спецификация работы*) устанавливает, какими качествами должен обладать работник для выполнения работы на должном уровне, т.е. его умения, знания, опыт, физические кондиции. Например, в Словаре "Наименования Профессий" Департамента труда США, наряду с более чем 20000 наименований профессий указываются и способности, которыми должен обладать работник для успешной работы в данной области. Каждое требование или характеристика обозначены буквами следующим образом:

- G (интеллигентность)

- V (словарный запас)
- N (численность)
- S (пространственное мышление)
- P (восприятие)
- Q (конторское восприятие)
- K (координация)
- F (ловкость рук)
- M (способность к руководству)
- E (координация “глаза-руки-ноги”)
- C (кругозор)

Ниже представлен список вопросов, используемых для составления спецификации работы [12]:

1. Название работы.

2. Общее изложение обязанностей.

3. Требуемый уровень образования:
   - базовое среднее,
   - послевузовское образование/профессиональная квалификация,
   - среднее специальное,
   - диплом по специальности.

4. Необходимость опыта аналогичной или родственной работы для лица, поступающего на работу:
   - не требуется никакого,
   - меньше 3 месяцев,
   - от 3 месяцев до 1 года,
   - от 1 до 3 лет.

5. Какова плотность контроля, требуемого на данной работе?
   - постоянный контроль,
   - несколько раз в день,
   - периодический,
   - ограниченный контроль,
   - минимальный или никакого.

6. Количество людей под началом работника (чел.):

- нет,
- 1,
- 2-5,
- 6-20,
- 21-50,
- 51 и более.

7. Во сколько обойдется компании ошибка, допущенная работником (тыс. долларов):

- меньше 50.0,
- 50.0-200,0,
- 200,0-1000,0,
- 1000.0-10000,0,
- больше 10000.0.

8. Как быстро может быть обнаружена ошибка?

- ежедневно,
- еженедельно,
- ежемесячно,
- ежеквартально,
- ежегодно,
- нет механизма регулярной проверки.

9. Контакты с другими людьми по инициативе работника:

- постоянно,
- часто,
- периодически,
- никогда.

10. Круг контактов:

- в собственном подразделении,
- в других подразделениях.
- с поставщиками,
- с заказчиками,
- с представителями власти,
- прочие.

11. Аспекты работы, требующие соблюдения конфиденциальности (секретности).

12. Опасные, вредные аспекты работы.

13. Требующаяся инициатива и изобретательность.

Выделив, в порядке их важности, конкретные задачи, решаемые на данном рабочем месте, аналитик затем для каждого задания указывает:

1. Требуемые знания (например, факты или принципы, с которыми исполнитель должен быть ознакомлен прежде, чем выполнять свою работу, процедуры);
2. Требуемые навыки (например, навыки, необходимые для управления машиной или другим транспортным средством);
3. Требуемые способности (например, математические, способность рассуждать, решать задачи, навыки вербального общения);
4. Физическая нагрузка (например, переноска тяжестей, необходимость таскать или толкать);
5. Все особенности рабочего места (вибрация, неадекватная вентиляция, перемещающиеся объекты или тесные помещения);
6. Типичные происшествия или инциденты на рабочем месте (например, напряженная работа в опасных условиях, работа с людьми);
7. Области интересов работника (предпочтение, которое работник должен отдавать работе с “вещами и объектами” или “передаче данных”, или “работе с людьми”).

На сегодняшний день существует значительное число методов анализа содержания работы:

1. Наблюдение (*Observation*).
2. Самоописание/дневник (*Self-description/diaries/logs*).
3. Метод критических случаев (*Critical incident technique*).
4. Интервью с целью анализа работы (*Job analysis interviews*).
5. Репертуарные решетки (*Repertory grid*).
6. Контрольные листы (*Checklists/inventories*).
7. Иерархический анализ заданий (*Hierarchical task analysis*).
8. Анализ обучения работе (*Job learning analysis*).
9. Исследование компонентов работы (*Job components inventory*).
10. Опросник позиционного анализа (*The position analysis questionnaire*).
11. Наблюдение участника (*Participant observation*).
12. Экспертные группы (*Expert conferences*).
13. Система опросов о производительности труда (*Work performance survey system*).

14. Комбинированный метод анализа работы (*Combination job analysis method: C-JAM*).
15. Ускоренный метод анализа работы (*Brief job analysis method: B-JAM*).
16. Функциональный анализ работы (*Functional job analysis*).
17. Метод выделения элементов работы (*The job element method*).
18. Проверка элементов работы (*Job element examining*).
19. Шкалирование востребованности возможностей (*Ability requirement scales*).

Как было отмечено выше, эффективный анализ содержания работы должен позволить выделить те аспекты профессиональной деятельности, которые позволяют четко различать хороших и плохих работников. Данный подход реализован в *методе критических случаев* (*инцидентов*).

- **Метод критических случаев**

Метод критических случаев (*critical incident technique*) был предложен в 1945 году Фланаганом (J.C. Flanagan) и относится к категории методов профессиографического анализа деятельности [11]. Этот метод связан со сбором сотен описаний эпизодов эффективных и неэффективных трудовых действий, которые реально наблюдали в своей трудовой деятельности опытные специалисты, руководители и другие работники. Эти эпизоды, названные "критическими случаями", должны представлять собой специфические действия, которые иллюстрируют успех или неудачу в одной из сторон анализируемого вида деятельности. Например, критическим неэффективным случаем для водителя грузовика может являться: "Водитель не посмотрел в зеркало заднего вида, когда давал задний ход, и в результате врезался в припаркованную машину". Метод применяется к группе работников, занимающих определенные должности, и/или их менеджеров или других «экспертов».

Наблюдателя, вспоминающего критический случай, как правило, просят описать:

1. Что послужило причиной данного случая и ситуацию, в которой он произошел?
2. В чем именно заключалась эффективность или неэффективность действий индивидуума?
3. Каковы очевидные последствия этих действий?
4. В состоянии ли был индивидуум контролировать эти последствия?

После того как набирается несколько сотен критических случаев, они подвергаются *контент-анализу*[12] и классифицируются одним или несколькими экспертами по категориям или "измерениям" критического рабочего поведения. Эти измерения впоследствии служат основой для проверки или разработки тестов и других процедур профессионального отбора. Их также можно использовать как базис при разработке программ профессионального обучения. Важное преимущество метода критических случаев как метода профессиографического анализа заключается в том, что он фокусируется на наблюдаемом и поддающемся измерению рабочем поведении. К недостаткам этого метода можно отнести то, что его реализация требует много времени и сил, а также его пренебрежение средним уровнем трудовой эффективности.

Приведем здесь последовательность действий анализа компетенций[13] методом критических случаев, описанную в своей книге М. Армстронгом [1]:

- Объясните респонденту, что представляет собой метод и для чего он используется, то есть скажите:

  «*Оцените, что составляет высокие и низкие показатели труда, анализируя события, которые, как было видно, имеют заметный успешный или неуспешный результат. Так вы предоставите больше фактической и «реальной» информации, чем просто перечисление задач и приблизительное определение требований к показателям труда*».

- Согласуйте и составьте список ключевых зон ответственности — основной подотчетности — в анализируемой работе. Чтобы сэкономить время, аналитик может определить их до встречи, однако необходимы гарантии того, что они были предварительно согласованы с группой, которой можно сказать, что список будет значительно улучшен в результате предстоящего анализа.

- Берите по очереди каждую область работы и просите группу привести примеры критических случаев. Если, например, одной из рабочих обязанностей является общение с покупателями, вопрос может быть задан следующим образом:

  "*Я хочу, чтобы вы рассказали мне о конкретных случаях в работе, связанной с общением с покупателями, в которых вы участвовали или которые вы наблюдали. Вспомните, в каких обстоятельствах они происходили, например, кто принимал*

[12] ***Контент-анализ*** (*content analysis*) - формализованный метод изучения текстовой и графической информации, заключающийся в переводе изучаемой информации в количественные показатели и ее статистической обработке.

[13] ***Компетенция*** – совокупность знаний, умений и аттитюдов (отношений, ценностей, мотивации), необходимых для эффективного выполнения конкретных задач, и которые исполнители демонстрируют в реальном поведении.

участие, что спрашивал покупатель, что вы или другой работник делали и что было в результате".

- Сгруппируйте информацию о критических случаях по следующим темам:
  - какие были обстоятельства;
  - что делал работник;
  - результат того, что делал работник.
  Эту информацию следует записать на плакатах, висящих на стене.
- Продолжайте этот процесс по каждой зоне ответственности.
- Обратитесь к плакатам и проанализируйте каждый случай, добиваясь того, чтобы участники выставили оценки приведенным примерам поведения по шкале от 1 по отношению к наименее эффективному поведению до 5 — для наиболее эффективного.
- Обсудите эти оценки, чтобы получить первоначальные определения эффективного и не эффективного выполнения для каждого ключевого аспекта работы.
- При необходимости доработайте эти определения после встречи — убедить группу выработать окончательные определения может быть затруднительно.
- Проведите окончательный анализ, в результате которого можно составить список требуемых компетенций и включить в него показатели или стандарты выполнения (поведения) для каждого принципиального подотчетного продукта деятельности или основной задачи.

Развитием метода критических случаев можно рассматривать *интервью с целью анализа работы* (*job analysis interviews*) или *структурированное интервью* (*structured interviews*).

- **Интервью с целью анализа работы**

Чтобы узнать все особенности работы, необходимо провести интервью с работниками и уточнить полученные данные у их менеджеров или руководителей групп. Целью интервью должно быть получение всех существенных данных, относящихся к работе.

Интервью для анализа работ М. Армстронг [1] рекомендует проводить следующим образом:

- Составить вопросы в логической последовательности, которая поможет работникам, с которыми проводится интервью, упорядочить свои представления о работе.

- Заранее выяснить необходимую информацию, чтобы определить, что работники делают: ответы на вопросы зачастую расплывчаты и могут предоставлять информацию в виде нетипичных примеров.

- Гарантировать, что работники не смогут отделаться общим или завышенным описанием своей работы — если, к примеру, интервью является частью оценки работы, будет странно, если работники не представят свою работу в самом лучшем свете.

- Отделить «зерна от плевел»: ответы на вопросы могут дать много несущественных данных, которые необходимо отсеять перед подготовкой должностной инструкции.

- Получить от работников ясное изложение их полномочий в принятии решений и количества указаний, которые они получают от менеджеров или руководителей групп. Это нелегко — если спросить, какие решения вы уполномочены принимать, большинство работников будут озадачены, потому что они думают о своей работе с точки зрения обязанностей и задач, а не абстрактных решений.

- Избегать наводящих вопросов, из которых ясно, какого ответа ожидают.

- Предоставить работнику возможность высказаться, создав атмосферу доверия.

Проводя интервью, полезно использовать перечень вопросов. Сложные перечни не нужны; они только вводят работников в замешательство. Суть искусства анализа работ «в его простоте». Стоит охватить следующие моменты [1]:

- Как называется ваша должность?
- Кому вы подчиняетесь?
- Кто вам подчиняется? (Полезно иметь структурную схему организации.)
- В чем состоит главная цель вашей работы? (То есть, в общих словах, чего от вас ожидают?)
- Что влияет на вас в процессе достижения этой цели? (Например, ответственность перед руководством, ключевые результаты или основные задачи.) Опишите, что вы должны делать, а не то, как вы это делаете. Также укажите, почему вы должны это делать, то есть те результаты, которых от вас ожидают.

- Какие параметры используются в вашей работе? (Используйте такие термины, как план производства или продаж, количество рассмотренных вопросов, количество подчиненных, количество клиентов.)
- Что еще вы можете рассказать о вашей работе в дополнение к сказанному, например:
  - насколько ваша работа сопряжена с другими работами в вашем отделе или в других подразделениях компании;
  - каковы требования к гибкости (разнообразны ли задачи, которые вы должны решать);
  - каким образом перед вами ставят задачи и каким образом рассматривают и утверждают вашу работу; ;
  - ваши полномочия в принятии решений;
  - с кем вы контактируете внутри и вне компании;
  - какое оборудование, механизмы и инструменты вы используете;
  - другие характерные особенности вашей работы, такие, как командировки, напряженность, требования к выносливости, работа в ночное или вечернее время, опасная работа;
    - с какими основными проблемами вы встречаетесь, выполняя вашу работу;
    - какие знания и навыки вам необходимы, чтобы выполнять свою работу.

Цель данного перечня вопросов — структурировать интервью при анализе работ в соответствии с вышеприведенными заголовками.

Если мы анализируем содержание работы с позиций компетенций, то на первом шаге «эксперты» составляют список задач или компетенций для интервьюирования. Структурированное интервью начинается с определения сферы ключевых результатов или основной подотчетности роли и продолжается анализом поведенческих характеристик, которые отличают работников с разным уровнем компетентности [1].

Основной вопрос:

«*Каковы положительные или отрицательные показатели поведения, которое приводит или не приводит к высокому уровню выполнения работы*?»

Примерный перечень анализируемых показателей поведения (компетенций), согласно М. Армстронгу [1], следующий:

- направленность личности (мотивация достижения);

- влияние на результаты;
- аналитические способности;
- стратегическое мышление;
- творческое мышление (способность вводить новации);
- настойчивость;
- коммерческий взгляд;
- руководство и лидерство;
- межличностные взаимоотношения;
- способность передавать информацию;
- способность адаптироваться и справляться с изменениями и напряжением;
- способность планировать и управлять проектами;
- склонность делиться знаниями.

Для каждой области (компетенции) эксперты пытаются определить примеры эффективного поведения.

Преимущества метода интервью состоят в том, что он очень гибкий, может предоставить всестороннюю информацию и его легко организовать и подготовить. Однако сам процесс интервью может отнимать много времени, а его результаты не всегда легко анализировать. В этом причина того, что в большинстве случаев при проведении анализа используются *контрольные* или *опросные листы*, которые дают предварительную информацию о работе, ускоряя этим процесс интервью или даже полностью его заменяя, хотя это средство может упустить многие «изюминки» работы, то есть того, что она представляет собой в реальности, а эти особенности необходимы для достижения максимально полного понимания роли работника.

Проблема структурированного интервью состоит в том, что оно слишком в большой мере опирается на способность эксперта получить информацию от опрашиваемых. Кроме того, нежелательно применение дедуктивного подхода, который, например, заменяет анализ подготовкой списка компетенций. Гораздо лучше использовать индуктивный подход, который начинается с определения типов поведения и затем группирует их по компетенциям. Это можно сделать в *экспертных группах* с помощью анализа положительных и отрицательных показателей, которые улучшают понимание компетенций в профессии или работе.

- **Контрольные листы**

Работники могут заполнять ***контрольные (опросные) листы*** (*checklists/inventories*), которые содержат вопросы, включенные в рассмотренный ранее перечень для проведения интервью, а их менеджеры или руководители групп могут подтвердить их. Опросные листы экономят время интервью, фиксируя исключительно связанную с фактами/фактическую информацию и давая аналитику возможность заранее составить вопросы так, чтобы охватить те области, которые требуют более глубокого исследования.

Преимущество опросных листов состоит в том, что они могут быстро и дешево давать информацию о большом количестве работ. Однако для них требуется большая выборка, а построение опросного листа — работа, которая требует квалификации и может быть выполнена только на основе некоторого предварительного сбора данных на местах. Перед началом полномасштабного использования опросного листа настоятельно рекомендуется провести его пробное испытание. Точность результатов также зависит от желания и способности заполнить опросный лист. Многим трудно выражать свое представление о работе в письменной форме, даже если они хорошо знают и выполняют ее.

Разновидностью контрольного листа может служить *контрольный перечень*, который дается работникам для заполнения. Он похож на опросный лист, но ответы требуют меньше субъективных суждений и предлагают варианты ДА или НЕТ. Перечни могут включать до 100 видов деятельности; работники отмечают галочкой те задачи, решение которых подразумевает их работа.

Как и опросные листы, перечни необходимо тщательно готовить, а важным является испытание на местах. Оно гарантирует, что инструкции по выполнению удовлетворительны и что ответы имеют смысл. Перечни можно использовать только при большом количестве работающих на данной должности. Если выборка меньше 30, результаты могут быть непредсказуемыми.

Оценочные шкалы представляют собой улучшенный вариант относительно грубого контрольного перечня. Как и перечень, они предлагают работникам список видов деятельности. Однако вместо простой просьбы отметить те из них, которые работники выполняют, шкалы предлагают оценить их (обычно от одного до семи), в соответствии с количеством времени, затрачиваемым на эту деятельность и иногда важностью задачи. Эти шкалы могут иметь такой вид, как показано в таблице 27.

Оценочная шкала для анализа работ		
Описание деятельности	Время, затрачиваемое на осуществление данной деятельности	Важность деятельности
Запросы информации по телефону	1. Почти нисколько (менее 10%) 2. Малая часть работы (10-24%) 3. Менее половины работы (25-44%) 4. Почти половина работы (45-54%) 5. Довольно большая часть работы (55-74%) 6. Очень большая часть работы (75-89%) 7. Почти вся работа (90% и более)	1. Совсем не важно 2. Не слишком важно 3. Не очень важно 4. Имеет некоторую важность 5. Важно 6. Очень важно 7. Крайне важно

Таблица 27. Пример оценочной шкалы для анализа работ [1].

Существует ряд доступных универсальных шкал, из которых самое широкое применение находит Опросный лист анализа должностей (Position Analysis Questionnaire), разработанный Мак-Кормиком и др. (McCormick et al, 1972). В его основе лежало исследование более 3700 работ, на основании чего были выделены шесть основных рабочих факторов:

- ввод информации;
- умственная деятельность; например, принятие решений;
- трудозатраты; например, использование механического контроля;
- взаимоотношения с людьми;
- рабочая среда;
- другие характеристики.

Для каждого заголовка, измеряющего специфические требования для почти 200 факторов работы, разработана своя шкала. Каждая шкала описывает определенную деятельность, и существует эталон для каждого пункта оценки.

Преимущества опросного (контрольного) листа для анализа должностей в широте применения, всесторонности и наличии эталонов. Однако его использование занимает много времени и требует некоторых специальных знаний.

- **Экспертные группы**

В ***экспертные группы*** (*expert conferences*) или ***рабочие группы*** (*working group*) входят люди, которые обладают «экспертными» знаниями или опытом работы, — менеджеры или работники — и посредник, обычно, но не обязательно, сотрудник отдела персонала или внешний консультант. Рабочая группа начинается с анализа «центральных» аспектов компетентности в данной организации: какие качества должны быть использованы в работе, чтобы достичь успеха (анализ содержания работы с позиций компетенций). Затем согласуются сферы рабочей компетентности — ключевые действия, которые выполняются работниками в рассматриваемой должности. Они определяются в терминах продуктов, то есть результата, который должен быть достигнут в конкретном аспекте данной должности.

Используя сферы компетенций в качестве основы, члены группы приводят примеры эффективного поведения, то есть поведения, которое, по всей вероятности, создает желаемые результаты [1]. Основной вопрос:

«*Что они делают и каким образом ведут себя, когда эффективно исполняют свою роль*?»

Ответы на этот вопрос будут высказываться в такой форме:

«*Человек в этой роли хорошо ее исполняет, когда он/она…*».

Примеры обсуждаемого типа поведения можно приводить отовсюду, откуда возможно. Ответы пишут на плакатах, висящих на стене. Далее группа, с помощью посредника, анализирует свои ответы и переводит их в ряд компетенций, которые определяются в терминах фактического поведения, указанного ранее. Слова, которые употребляла группа, используются максимально для того, чтобы они могли «владеть» результатом. Эти компетенции формируют основание для структуры общих или профиля специфических компетенций.

Например, одной из сфер компетенции роли директора по управлению персоналом может быть планирование человеческих ресурсов, определенное как:

> "*Составление прогнозов потребностей в человеческих ресурсах и планов их приобретения, сохранения и эффективного использования, которые гарантируют, что потребности компании в человеческих ресурсах удовлетворены*".

Аспекты компетенции в этой сфере можно выразить как: «*Некто в этой должности будет хорошо ее исполнять, если он/она*»:

- добивается участия в разработке стратегии коммерческой деятельности;
- вносит свой вклад в бизнес-планирование, придерживаясь стратегической точки зрения на перспективные вопросы, связанные с человеческими ресурсами, которые могут повлиять на стратегию коммерческой деятельности;
- налаживает связи с коллегами из высшего руководства, чтобы понимать вопросы планирования человеческих ресурсов, которые они ставят, и отвечать на них;
- предлагает практические способы совершенствования использования человеческих ресурсов.

Роль посредника в рабочей группе состоит в том, чтобы стимулировать группу, помогать ей анализировать свои находки и, в целом, способствовать созданию набора компетенций, которые можно иллюстрировать примерами поведения.

Еще одним методом, отличающим высокие стандарты выполнения от низких, может служить методика репертуарных решеток.

- **Репертуарные решетки**

Репертуарные решетки (*Repertory grid*) основаны на теории личностных конструктов Дж. Келли (Kelly, 1955). Личностные конструкты представляют собой способы, которыми мы смотрим на мир. Они личностные, потому что в значительной степени индивидуальны и влияют на наше поведение и на наше мнение о поведении других людей [1].

Аспекты работы, к которым применяются эти «конструкты» или суждения, называются «элементами». Чтобы узнать эти суждения, группу людей просят сосредоточиться на определенных элементах, которые представляют собой задания, выполняемые работниками, и выработать конструкты относительно этих элементов. Это дает им возможность определить, что указывает на существенные требования к успешному выполнению.

Процедура, которой придерживается аналитик, известна как «триадный метод вывода» (разновидность фокуса с тремя картами) и включает в себя, согласно М. Армстронгу [1], следующие шаги:

1. Определите, какие задания или элементы работы анализируются методом репертуарных решеток. Это делается с помощью одной из форм анализа работ, например интервью.
2. Перечислите задания на карточках.
3. В случайном порядке вытяните из колоды три карты и попросите членов группы назвать то из этих заданий, которое превосходит другие с точки зрения качеств и характеристик, необходимых для его выполнения.
4. Попробуйте получить более конкретные определения этих качеств или характеристик в терминах ожидаемого поведения. Если характеристики описываются, к примеру, как «способность планировать и организовывать», задайте такой вопрос, как: «Какое поведение или какие действия показывают, что кто-то планирует эффективно?» или «Что мы можем сказать, если кто-то не особенно хорошо организует свою работу?»
5. Вытащите из колоды еще три карты и повторите шаги 3 и 4.
6. Повторяйте этот процесс до тех пор, пока все карточки не будут проанализированы и больше не останется конструктов, которые нужно определить.
7. Составьте список конструктов и попросите членов группы оценить каждое задание по каждому качеству, используя шести или семибалльную шкалу.
8. Подсчитайте и проанализируйте количество набранных очков, чтобы оценить их относительную важность. Это можно сделать статистически.

Как и метод критических случаев, метод репертуарных решеток помогает работникам четко высказать свое мнение относительно конкретных примеров. Дополнительным преимуществом является то, что репертуарные решетки облегчают работникам процесс определения поведенческих характеристик компетенций, требуемых в работе, ограничивая область сравнения с помощью триадного метода.

Хотя полный статистический анализ результатов, полученных методом репертуарных решеток, полезен, самые важные результаты, которые можно получить, представляют собой описания того, из чего состоит хорошее и плохое выполнение в каждом элементе работы.

Как репертуарные решетки, так и метод критических случаев, требуют участия квалифицированного специалиста, который может исследовать и составить описания характеристик работы. Они очень подробны и требуют времени, но даже если полностью

процесс не проводится, большая часть его методологии полезна для менее сложного подхода анализа работы (или компетенций).

Сосредоточение на критических требованиях конкретных видов профессиональной деятельности привело к разработке ***метода рабочих элементов*** (*job element method*) для конструирования тестов и доказательства их содержательной валидности. В нем *рабочие элементы* - это единицы описания критических требований, предъявляемых конкретным видом работы к работнику. Описание профессиональной деятельности на языке специфических требований к поведению работника позволяют в дальнейшем прямо формулировать задания теста. Конкретные поведенческие формулировки могут, в свою очередь, объединяться в более широкие категории, или конструкты, - такие как точность вычислений, развитая тонкая моторика, зрительное различение или способность работать под давлением *(to work under pressure).* Данный метод слабо представлен в отечественной литературе.

Кратко остановимся еще на ряде методов анализа содержания работы.

- **Иерархический анализ заданий**

Иерархический анализ заданий (*hierarchical task analysis*) или *анализ иерархии задач*, разработанный Д. Аннетом и К. Дунканом (Annet and Duncan, 1971), разбивает работы или их сферы на задачи, подзадачи и планы, расположенные в иерархическом порядке. Задачи определяются в терминах целей или конечных продуктов, и планы достижения целей также подвергаются анализу. Процесс начинается с анализа общей задачи. Затем строится иерархия подзадач вместе с их продуктами и определяются промежуточные планы их достижения.

Этот метод предполагает [1]:

- использование активной формы глаголов, которые понятно и конкретно описывают то, что нужно делать;
- определение стандартов выполнения работы, то есть, уровня показателей труда, которые должны быть достигнуты, для того чтобы считать задачу или операцию выполненной удовлетворительно;
- перечисление условий, связанных с показателями выполнения задачи, которые могут включать в себя факторы среды, такие, как работа в зоне высокого шума.

Обычно этот подход применяется к работам, связанным с переработкой или производством, но принципы анализа подзадач и определения стандартов продукции и показателей труда подходят для анализа работ любого типа.

- **Наблюдение**

Наблюдать — означает изучать работников за работой, подмечая, что они делают, как они это делают и сколько времени на это уходит. Метод ***наблюдения*** (*observation*) можно использовать, смотря по ситуации, там, где относительно малое количество ключевых работ и необходим глубокий анализ. Но его применение требует много времени, и его трудно осуществить в отношении работ, подразумевающих в основном недоступную для наблюдения умственную деятельность или высококвалифицированный ручной труд, в котором действия выполняются слишком быстро, чтобы безошибочно их замечать.

- **Самоописание**

Можно попросить работников проанализировать собственную работу и подготовить ее описание (***саомоописание*** - *self-description*). Это значительно экономит аналитику время, которое он может потратить на проведение интервью или наблюдение за работником. Однако работникам не всегда легко это сделать, возможно, потому, что то, что они делают, в значительной мере составляет часть их самих, и им трудно разделить информацию на различные элементы. Поэтому в большинстве случаев требуется некоторое руководство. Если в опрос вовлечен целый ряд работников (например, они должны оценить работы), рекомендуется провести специальные учебные занятия, на которых они смогут практиковаться в анализе своей собственной работы и работы других людей.

- **Дневники и рабочие журналы**

Этот подход (***дневники и рабочие журналы*** - *diaries/logs*) требует от работников, чтобы они анализировали свою работу, отмечая свои действия в дневниках или рабочих журналах. Специалист по анализу работ может взять их за основу для составления, например, должностных инструкций. Работникам необходимо разъяснить, каким образом вести дневник или рабочий журнал. Их можно попросить описывать обычный день час за часом или записать свои действия в форме рассказа по окончании какого-либо периода времени, обычно дня. Дневники и рабочие журналы лучше всего использовать в достаточно сложной управленческой деятельности и там, где у работников есть необходимые аналитические навыки и способность выражать свои мысли в письменной форме.

- **Общие процедуры анализа содержания работы**

Общая процедура анализа содержания работы многоступенчата и включает в себя следующие этапы (см. табл. 28).

Основные этапы	Содержание
1. Анализ работы (Свойства работы)	• Описание содержания работы (что должно быть сделано – перечень задач) и характеристика ее выполнения (как должно быть сделано - результаты). • Выявление условий труда. ***Проблема:*** разграничение элементов деятельности.
2. Анализ требований к исполнителю (Свойства работников)	• "Перевод" характеристик работы в характеристики работников, т.е. в "квалификации" (знания, умения, опыт). • "Предпосылки пригодности" – психофизиологические требования работы. ***Проблема:*** распознавание на основе видимого (задачи, работы) незримого (психофизиологических характеристик работника).
3. Анализ значимости	• В какой степени требуется выполнение отдельных предпосылок пригодности; насколько "значимы" отдельные требования. ***Проблема:*** Критерии значимости; индивидуальные отличия в способностях.
4. Разработка профиля требований к рабочему месту (должности)	• Определение всех предпосылок, которым должен отвечать человек, чтобы успешно справиться с соответствующими задачами. Основу профильного метода составляет каталог характеристик-требований, предъявляемых к человеку в зависимости от

Основные этапы	Содержание
	выполняемой им работы, а также с учетом количественных характеристик рабочих мест и персонала.

Таблица 28. Этапы анализа содержания работы [12].

Федеральным правительство США был разработан следующий поэтапный подход к проведению типичного анализа содержания работы, который представлен в таблице 29:

Шаги	Что нужно запомнить или сделать
1. Собрать пред-варительную информацию о работе	• Просмотреть существующие документы, чтобы сконструировать «панорамное» представление о работе: ее главное назначение, основные обязанности или функции, схемы трудового процесса. • Подготовить предварительный список обязанностей, который будет служить рамками для проведения интервью. • Отметить те основные положения, в которых нет ясности, или есть двоякое толкование, или их нужно уточнить в процессе сбора данных.
2. Провести начальный осмотр рабочего места	• Начальный осмотр введен для ознакомления аналитика с организацией деятельности, используемыми инструментами и оборудованием, общими условиями труда и механикой, сопровождающей сквозное выполнение основных обязанностей. • Начальный осмотр особенно полезен в тех работах, где лучше один раз увидеть сложное и незнакомое оборудование, чем заставлять человека, отвечающего на вопросы, тратить тысячу слов на описание незнакомого вам предмета или техники. • Для сохранения последовательности рекомендуется, чтобы сопровождающим для осмотра рабочего места назначили интервьюируемого руководителя первого уровня[14].

[14] Руководитель первого уровня является непосредственным руководителем сотрудника.

Шаги	Что нужно запомнить или сделать
3. Провести интервью	• Рекомендуется проводить первое интервью с руководителем первого уровня, позиция которого считается более выгодной по сравнению с позицией исполняющего обязанности, для получения общего представления о работе и о том, как сочетаются основные обязанности. • С целью разработки графика рекомендуется приводить не больше двух интервью в день, каждое не более 3 часов.
Замечания по отбору интервьюируемых	• Интервьюируемые считаются знатоками своего дела на основании того факта, что они выполняют данную работу (в случае тех, на кого возложены те или иные обязанности) или отвечают за то, чтобы она была сделана (в случае руководителей первого уровня). • Приглашаемый на интервью вступивший в должность сотрудник должен быть типичным служащим, который хорошо осведомлен о работе (не ученик, только начинающий ориентироваться или выдающийся представитель рабочего подразделения). • Где только возможно, люди для интервью должны приглашаться так, чтобы получить подходящее сочетание по расовым признакам и полу.
4. Провести второй осмотр рабочего места	• Второй осмотр рабочей площадка предназначен для уточнения, подтверждения и других видов обработки информации, набранной из интервью. • Как и в начальном осмотре, рекомендуется, чтобы тот же интервьюируемый руководитель первого уровня проводил второй обход.

Шаги	Что нужно запомнить или сделать
5. Свести всю информацию о работе	• Консолидирующий этап изучения работы - это сведение в одно последовательное и полное описание работы всех данных, полученных из разных источником: от руководителя, государственных служащих, от посещения площадки и из письменных материалов о работе. • Опыт прошлых лет показывает, что на каждую минуту интервью требуется минута подведения итогов. Исходя из интересов планирования, следует оставить, по крайней мере, пять часов на консолидирующий этап. • На этапе консолидации аналитик должен иметь постоянную возможность общаться со знатоком темы как источником сведений. Эта роль отведена интервьюируемому руководителю. • Аналитик должен свериться с начальным предварительным списком обязанностей и вопросов - все вопросы должны быть сняты.
6. Проверить описание работы	• На этапе проверки все опрошенные собираются вместе, с целью определения, что сводное описание работы получилось точным и полным. • Процесс проверки идет в группах. Руководителю первого уровня и исполняющим должностные обязанности лицам, прошедшим интервью, раздают отпечатанные или разборчиво написанные копии описаний работы (повествовательное описание рабочей обстановки и список формулировок заданий). • Строчка за строчкой аналитик проходит по всему описанию работы и отмечает любые упущения, неясности или места, нуждающиеся в уточнении. • В конце проверочного собрания аналитик собирает все материалы.

Таблица 29. Общий порядок проведения типичного анализа содержания работы, разработанный Федеральным правительство США [22].

В следующей таблице представлены типичные данные, собранные для анализа содержания работы (см. табл. 30):

<table>
<tr><th colspan="3">Данные, связанные с работой</th></tr>
<tr><td colspan="2">Установление работы
Должность
Отдел, в котором выполняется эта работа
Число людей, исполняющих работу</td><td>Содержание работы
Задания
Занятия
Ограничения действий
Критерии эффективности
Случаи, подлежащие критике
Противоречащие требования
Условия работы
Роли (например, ведущий переговоры, наставник, лидер)</td></tr>
<tr><th colspan="3">Данные, связанные с сотрудником</th></tr>
<tr><td>Характеристика сотрудника
Профессиональные/технические знания
Умения ручного труда
Умения высказывать мысли
Умения письменно выражать мысли
Количественные умения
Механические умения
Концептуальные умения
Административные умения
Умение быть лидером
Навыки межличностного общения</td><td>Внутренние отношения
Босс и другие начальники
Коллеги
Подчиненные</td><td>Внешние отношения
Поставщики
Клиенты
Проверяющие организации
Профессиональная отрасль
Сообщество
Профсоюз/группы сотрудников</td></tr>
</table>

Таблица 30. Типичные данные, собранные для анализа содержания работы [22].

После определения заданий и результатов аналитики переходят к рассмотрению типов поведения, ведущие к этим результатам (характеристики сотрудника в таблице 29). В

таблицах 31 и 32 представлены примеры, как коммуникация может быть описана глаголами действия.

1. Отметьте кружок в графе "Делаете это" для заданий, которые вы выполняете в данное время. 2. В конце списка заданий впишите любые не перечисленные задания, которые в данное время исполняете. 3. Оцените каждое задание, которое вы выполняете относительно количества времени, отметив соответствующий кружок в графе "Время, потраченное на текущее занятие". Пожалуйста, используйте мягкий карандаш и полностью закрашивайте кружки.	Делаете это	Время, потраченное на текущее задание								
		1.Очень мало	**2**. Намного меньше среднего	**3**. Меньше среднего	**4**. Немного ниже среднего	**5**. Среднее	**6**. Немного больше среднего	**7**. Больше среднего	**8**. Намного больше среднего	**9**. Значительно больше среднего
Выполняете коммуникативные действия ***Получаете техническую информацию:***										
421. Читаете техническую литературу о конкретной продукции.	☐	1	2	3	4	5	6	7	8	9
422. Читаете техническую литературу, чтобы не отставать от развития отрасли.	☐	1	2	3	4	5	6	7	8	9
423. Посещаете требуемые, рекомендуемые или относящиеся к работе курсы и/или семинары.	☐	1	2	3	4	5	6	7	8	9
424. Изучаете существующие операционные системы/программы, чтобы научиться и продолжать пользоваться ими.	☐	1	2	3	4	5	6	7	8	9
425. Занимаетесь поиском литературы, необходимой для развития продукции.	☐	1	2	3	4	5	6	7	8	9
426. Общаетесь с группой системного программного обеспечения, чтобы понимать, как производимые перемены влияют на текущие проекты.	☐	1	2	3	4	5	6	7	8	9
427. Изучаете и даете оценку внедренным техническим приемам, чтобы оставаться конкурентоспособными и/или ведущими в своей области.	☐	1	2	3	4	5	6	7	8	9
428. Посещаете совещания по промышленным стандартам.	☐	1	2	3	4	5	6	7	8	9

Обмениваетесь технической информацией:	
429. Координировано работаете с кодировщиками, чтобы проверять, что проектирование программных средств осуществляется в установленном порядке.	☐ 1 2 3 4 5 6 7 8 9
430. Обращаетесь к коллегам по работе, чтобы обмениваться идеями и техническими приемами.	☐ 1 2 3 4 5 6 7 8 9
431. Обращаетесь к членам других технических групп внутри компании, чтобы обмениваться идеями и техническими приемами.	☐ 1 2 3 4 5 6 7 8 9
432. Координировано работаете с консультантами или организациями по поддержке, чтобы уточнять проектирование программных средств или содержание обучающих программ.	☐ 1 2 3 4 5 6 7 8 9

Таблица 31. Коммуникация: характеристики поведения [22].

Раздел 4. Отношения с другими	**Код важности для данной работы**
Этот раздел посвящен разным сторонам взаимодействия между людьми, участвующими в разных видах деятельности.	N. Не применяется 1. Очень редко 2. Редко 3. Средне 4. Часто 5. Чрезвычайно часто
4.1. Коммуникация	
Оцените следующие пункты сточки зрения важности этого занятия для завершения работы. Некоторые работы могут охватывать несколько или все пункты данного раздела.	
4.1.1. Устная (речевая коммуникация)	
99_____	Советовать (общаться с людьми для того, чтобы проконсультировать их и/или указать им путь к решению тех проблем, которые могут быть решены юридически, финансово, научно, технически, клинически, духовно и/или профессионально).
100_____	Вести переговоры (общаться с людьми для того, чтобы достичь соглашений по какому-то решению, например трудовая сделка, дипломатические отношения и т. д.).
101_____	Убеждать (общаться с людьми для того, чтобы подвести их к какому-то действию или точке зрения, например продаже, политической компании и т. д.).
102_____	Наставлять (обучать других знаниям или умениям, по-дружески или в официальной манере, например учитель средней школы, машинист, обучающий ученика и т д.).

103_____	Брать интервью (проводить интервью, ведущие к какой-то определенной цели, например проводить собеседования с претендентами на рабочее место, проводить перепись населения).
104_____	Осуществлять повседневный обмен информацией, связанный с работой (давать и/или получать информацию по работе повседневного характера, например диспетчер такси, служащий в приемной и т. д.).
105_____	Неповседневный обмен информацией (давать и/или получать информацию по работе неповседневного или необычного характера, например заседания профессиональных комиссий; инженеры, обсуждающие проект новой продукции и т. д.)
106_____	Общественное выступление (произносить речь или вести официальную презентацию перед относительно большой аудиторией, например политические общения, теле-, радиовещание, читать проповедь и т. д.).
4.1.2. Письменная (коммуникация посредством написанных/напечатанных материалов)	
107_____	Писать (например, писать или диктовать письма, отчеты и т.д., писать образец объявления, газетные статьи и т.д.; не включает стенографическую деятельность, описанную в пункте 4.3., а только ту деятельность, в которой государственный служащий сам сочиняет письменный материал).

Таблица 32. Коммуникация: характеристики поведения [22].

6. ТАБЛИЦЫ КРИТИЧЕСКИХ ЗНАЧЕНИЙ КОЭФФИЦИЕНТА КОНКОРДАЦИИ *W*

Эксперты	**Объекты (*n*)**					**Дополнительные значения *S* для *n* = 3 (*α* = 0.05)**	
m	**3**	**4**	**5**	**6**	**7**	***m***	***W***
3	--	--	0,7156	0,6597	0,6242	**9**	0,3333
4	--	0,6188	0,5525	0,5118	0,4844	**12**	0,2497
5	--	0,5008	0,4492	0,4169	0,3946	**14**	0,2393
6	--	0,4206	0,3781	0,3514	0,3325	**16**	0,1871
8	0,3758	0,3178	0,2870	0,2670	0,2528	**18**	0,1662
10	0,3000	0,2556	0,2312	0,2153	0,2039		
15	0,1996	0,1715	0,1555	0,1449	0,1373		
20	0,1496	0,1290	0,1171	0,1092	0,1035		

Таблица 33. Критические значения *W* для уровня значимости (*α*) = 0.05 [23].

Эксперты	**Объекты (*n*)**					**Дополнительные значения *S* для *n* = 3 (*α* = 0.01)**	
m	**3**	**4**	**5**	**6**	**7**	***m***	***W***
3	--	--	0,8400	0,7797	0,7365	**9**	0,4685
4	--	0,7675	0,6831	0,6293	0,5915	**12**	0,3594
5	--	0,6440	0,5712	0,5243	0,4911	**14**	0,3110
6	--	0,5528	0,4892	0,4483	0,4192	**16**	0,2738
8	0,5219	0,4294	0,3792	0,3467	0,3236	**18**	0,2448
10	0,4255	0,3506	0,3091	0,2823	0,2632		
15	0,2911	0,2398	0,2112	0,1926	0,1793		
20	0,2213	0,1821	0,1603	0,1460	0,1359		

Таблица 34. Критические значения *W* для уровня значимости (*α*) = 0.01 [23].