Методика разработки теста

Глава 1

Базовые определения

Дидактическая единица (ДЕ) -

- законченный по смыслу элемент знаний или умений, подлежащий усвоению в соответствии с ГОС (рабочей программой).

Детальность ДЕ должна быть различна для разных уровней контроля:

для итоговой аттестации могут быть приняты **разделы** дисциплин, приведенные в ГОС ВПО;

для текущего и промежуточного контроля - ориентация на **ДЕ рабочей программы** (элементы содержания описаны подробнее и точнее, чем в ГОС).

Педагогический тест -

В.С. Аванесов

- ***система заданий*** возрастающей трудности, ***специфической формы***, позволяющая качественно и эффективно ***измерить уровень*** и оценить структуру ***подготовленности учащихся***.

Комментарий № 1 к определению:

Система означает, что в тесте собраны такие задания, которые образуют систему, т.е. связаны между собой. В педагогическом тесте задания должны располагаться по мере возрастания трудности - от самого легкого до самого трудного. Т.е. главным системообра-зующим признаком теста является **различие заданий по степени их трудности**.

Комментарий № 2 к определению:

Специфическая форма тестовых заданий отличается тем, что задания теста представляют собой не вопросы и не задачи, а задания, сформулиро-ванные в форме высказываний, истинных или ложных, в зависимости от ответов.

Комментарий № 3 к определению:

Оценка уровня подготовленности означает использование в тесте только такого контрольного материала, который соответствует содержанию учебной дисциплины (но не проверку психологических качеств, например, развития мышления)

Достоинства тестового контроля:

- высокая **объективность** оценки уровня учебных достижений обучаемого за счет: **а)** стандартизации процедуры опроса; **б)** опроса по всему учебному материалу; **в)** применения статистических методов обработки результатов тестирования;
- тестирование – **педагогическая технология** (гарантированное достижение результата, переносимость, возможность автоматизации процедуры опроса);
- оперативность;
- массовость;
- возможность совершенствование теста за счет количественной оценки **валидности** и **надежности** результатов тестирования.

Типы тестов (по целям):

- **критериально-ориентированные** – выявление уровня усвоения ДЕ, выделенных в ГОС или учебной программе;
- **нормативно-ориентированные** – дифференциация (ранжирование) испытуе-мых либо по отношению к *норме*, либо по отношению *друг к другу*. Норма определяется статистическими методами по репрезента-тивной выборке протестированных (обычно основана на нормальном распределении);
- **аттестационные** педагогические измерительные материалы (АПИМ).

Основная задача тестирования –

- ***дифференциация*** обучаемых по ***уровню подготовленности*** (усвоения учебного материала).

Следствия (предмет размышлений):

1. Создание теста должно начинаться с определения **уровня контроля** и **цели тестирования** – именно они определяют содержание теста.
2. Для каждого уровня контроля требуют определения **принципы отбора содержания** теста.
3. Тест **не может** состоять из заданий одного уровня трудности, следовательно, требует определения процедура **разделения** тестовых заданий **по трудности**.

Следствия (предмет размышлений):

4. Тестовые измерительные материалы для контроля более высокого уровня **не могут быть** получены механическим объединением заданий из тестов более низкого уровня (поскольку в тест должны входить вопросы, касающиеся всей темы (раздела), требующие обобщения, сопоставления и пр. умственных действий высокого уровня).
5. Тест, который выполняется полностью большинством обучаемых или почти никем из них, **не имеет смысла**, поскольку не обеспечивает дифференциации. Следовательно, **должны быть установлены количественные характеристики качества теста.**

Уровни контроля:

a) вступительный (входной):

цель – установление соответствия знаний поступающего (начинающего изучать дисциплину) входным требованиям (специальности, дисциплины);

объем – в соответствии с программой вступительных испытаний или требованиями преемственности обучения;

b) оперативный (текущий):

цель – установление уровня усвоения текущего материала для обеспечения возможности коррекции хода учебного процесса;

объем – материал 1-2 занятий;

Уровни контроля:

c) **тематический**; проводится по завершении изучения темы;

цель – соотнесение результатов освоении ДЕ темы с требованиями учебной программы;

объем – материал темы (*все* ДЕ темы);

c) **рубежный**; проводится по завершении изучения раздела;

цель – соотнесение результатов освоении дидактических единиц раздела с требованиями учебной программы;

объем – материал раздела (*базовые* ДЕ раздела);

d) **итоговая аттестация**; проводится по завершении изучения дисциплины;

цель – установление уровня и качества подготовки обучаемых требованиям ГОС;

объем – материал дисциплины (*базовые* ДЕ ГОС).

Следствия:

1. Уровню контроля однозначно соответствует цель контроля и объем проверяемого материала .
2. С увеличением уровня контроля возрастает объем проверяемого материала и сложность умственных действий, необходимых для его выполнения.
3. Поскольку содержание теста должно соответствовать уровню контроля, цель тестирования должна быть определена изначально (до разработки содержания ТЗ).

Следствия:

4. Для каждого уровня контроля должны быть определены принципы отбора содержания теста.
5. Тест более высокого уровня не может быть построен механическим объединением вопросов из тестов более низкого уровня, поскольку в тест должны входить вопросы, касающиеся всей темы (раздела), требующие обобщения, сопоставления и пр. умственных действий высокого уровня.

Глава 2

Определение нормы трудности тестового задания

Понятия:

- **Концепт** – умственное действие (*умозаключение, знание формулы, правила, аксиомы, сопоставление и т.д.*), которое требуется для выполнения ТЗ.

- **Дистрактор –** предлагаемый в ТЗ закрытого типа альтернативный вариант ответа.

Норма трудности ТЗ:

– характеристика задания, устанавливаемая с помощью количественной порядковой шкалы в соответствии с оговоренными критериями (например, на основе экспертных оценок).

Норма трудности (НТ) учитывается при разработке структуры (спецификации) теста, а также при выставлении оценки за тест (в последнем случае НТ выступает в качестве весового множителя при подсчете средней доли выполнения задания).

Норма трудности определяется:

- *формой* (типом) *ТЗ* – в порядке убывания НТ: установление последовательности или соответствия, открытого типа; множественный выбор, одиночный выбор;
- *количеством дистракторов* – с ростом числа дистракторов НТ возрастает (поскольку требуется большее число концептов для выбора верных ответов).

Норма трудности определяется:

- *количеством* используемых *концептов*, необходимых для поиска правильного решения;
- *сложностью концептов* по типу мыслительных операций в соответствии с некоторой таксономией.

Таксономия Блума:

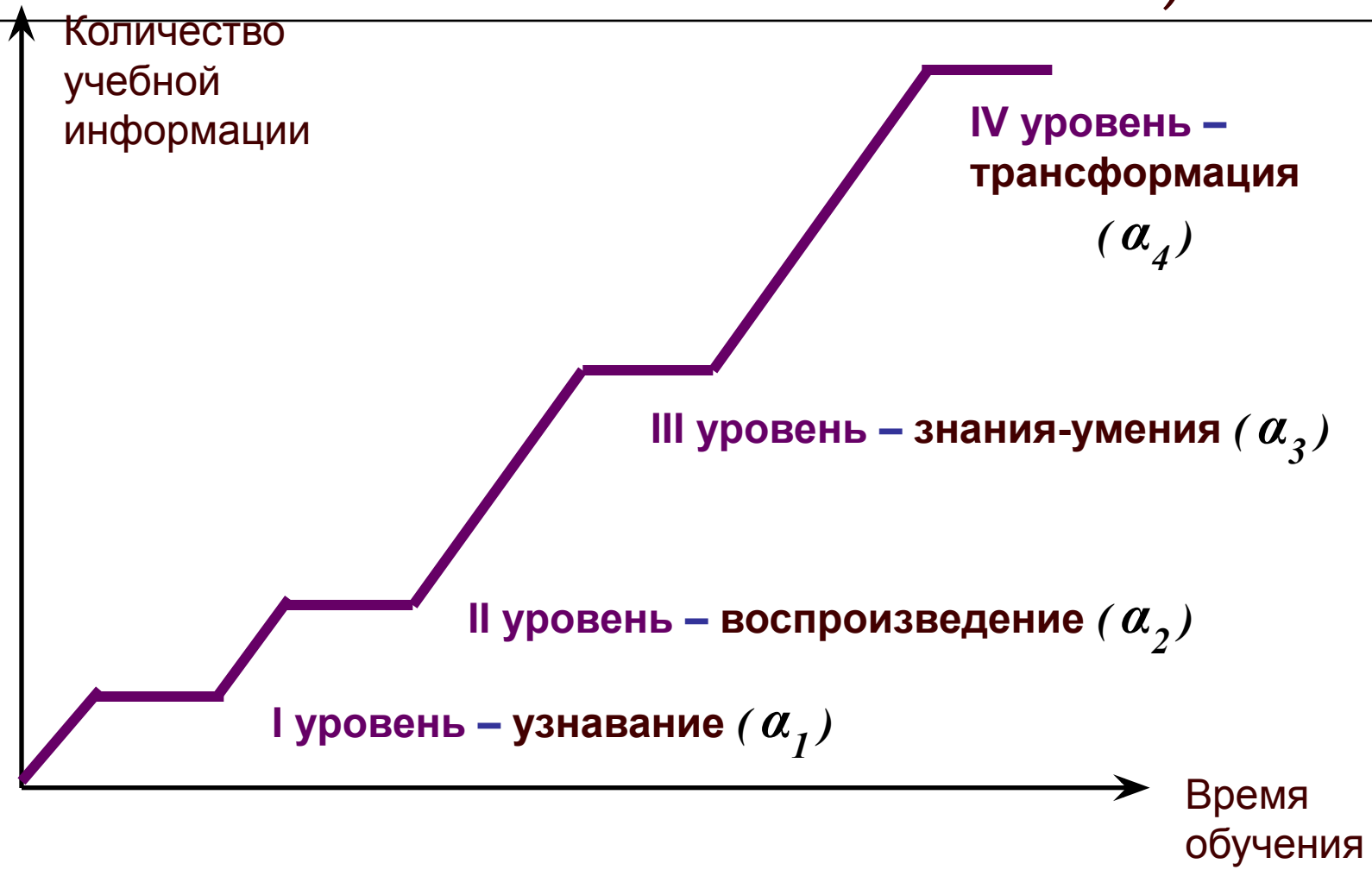
1. ЗНАНИЕ
    - 1.1. конкретного материала
        - 1.1.1. терминологии
        - 1.1.2. фактов
    - 1.2. способов и средств обращения с конкретным материалом
        - 1.2.1. конвенций (определений)
        - 1.2.2. тенденций и результатов
        - 1.2.3. системы понятий и категорий
        - 1.2.4. критериев
        - 1.2.5. методологии
    - 1.3. универсальных понятий и абстракций данной области знаний
        - 1.3.1. законов и обобщений
        - 1.3.2. теорий и структур
2. ПОНИМАНИЕ
    - 2.1. Объяснение
    - 2.2. Интерпретация
    - 2.3. Экстраполяция

Таксономия Блума:

3. ПРИМЕНЕНИЕ
4. АНАЛИЗ
    4.1. элементов
    4.2. взаимосвязей
    4.3. принципов построения
5. СИНТЕЗ
    5.1. Единичное сообщение
    5.2. Разработка плана и возможной системы действий
    5.3. Получение системы абстрактных отношений
6. ОЦЕНКА
    6.1. Суждение на основе имеющихся данных
    6.2. Суждение на основе внешних критериев

Процедура назначения НТ

(на основе подхода В.П. Беспалько)



Модель четырех уровней усвоения по В.П. Беспалько:

I уровень («*знания-знакомства*») – узнавание объектов, свойств, процессов данной области явлений действительности при повторном восприятии ранее усвоенной информации о них или действий с ними;

II уровень («*знания-копии*») – репродуктивное действие путем самостоятельного воспроизведения и применения информации по ранее усвоенной ориентировочной основе для выполнения известного действия;

III уровень («*знания-умения*») – продуктивное действие – деятельность по образцу;

IV уровень («*знания-трансформации*») – творческое действие, выполняемое путем самостоятельного конструирования новой ориентировочной основы для деятельности

Процедура назначения НТ (*подход Центра тестирования Васильева*)

Простое задание:

- ТЗ направлено на «опознание» какого-то объекта или на проверку «знания-знакомства»;
- ТЗ направлено на выбор одного варианта ответа из многих с помощью знания всего одного концепта;
- ТЗ открытого типа направлено на выявление знания определения односложного базового термина.

Процедура назначения НТ
(*подход Центра тестирования Васильева*)

Среднего уровня трудности:

- ТЗ направлено на применение усвоенных ранее знаний в типовых ситуациях (т.е. в тех ситуациях с которыми знаком испытуемый) или на проверку «знаний воспроизведения копии»;
- ТЗ, связанные с высказывания конъюнктивного или дизъюнктивного вида (сопоставление, объединение, вычленение и т.п.);
- или ТЗ с несколькими концептами по выбору подмножества правильных вариантов из заданного множества дистракторов (задания множественного выбора, на установления соответствия);
- в некоторых случаях – ТЗ на установление последовательности (порядка).

Процедура назначения НТ (*подход Центра тестирования Васильева*)

Сложные задания:

- ТЗ направлено на применение усвоенных ранее знаний и умений в нестандартных условиях (т.е. в условиях, ранее не знакомых испытуемому) или на проверку «знаний умения и применения»;
- ТЗ, сформулированные в виде утверждений *импликативного* типа (такие задания требуют применения рассуждений в форме дедуктивного, индуктивного вывода и аналогии, причем для получения окончательного ответа необходима некоторая последовательность умозаключения (несколько концептов)).

Вывод:

Возможны различные (иные) подходы к назначению нормы трудности ТЗ. ***Важно***, что выбранный подход должен оставаться ***неизменным*** в пределах данного теста.

Из практики: достаточно дифференцировать задания по ***трем*** уровням трудности – (обозначим их α_1, α_2 и α_3).

Глава 3

Этапы разработки теста

Этап 1. *Определение уровня контроля и цели тестирования*

Для преподавателя школы это могут быть уровни контроля текущего, тематического, рубежного.

Для преподавателя вуза – все виды контроля.

На данном этапе указывается также *контингент*, на который рассчитан тест. Возможно указание *учебной программы*, по которой шло обучение, или *учебника* (если их допускается несколько).

Возможные цели тестирования:

- проверка качества усвоения материала;
- ранжирование обучаемых по уровням успешности освоения дисциплины;
- итоговая аттестация;
- проверка соответствия знаний обучаемых требованиям ГОС;

… и др.

Этап 2. *Определение времени тестирования и числа заданий в тесте*

Поскольку на выполнение ТЗ отводится в среднем 1-3 мин., общее время тестирование определяет количество ТЗ в тесте.

Пример 1. Тест для текущего контроля, на который отведено 12-15 мин.: N = 6-7.

Пример 2. Тест для рубежного контроля, на который отведено 80 мин.: N = 40-50.

***Важно*: при фиксированном времени тестирования число ТЗ в тесте, практически, не зависит от уровня контроля.**

Этап 3. *Отбор содержания теста*

В тестах для ***текущего*** контроля содержание теста разработчик определяет в зависимости от содержания учебной программы, материала учебника, особенностей учебной группы и пр.

Проверке подлежат ***все*** ДЕ, которые осваивались на занятиях.

На начальных этапах обучения (освоение теории) это могут быть задания с нормой трудности α_1 и α_2. При отработке умений выполнения типичных заданий – α_2 и α_3.

Этап 3. *Отбор содержания теста*

В тестах для контроля тематического, рубежного, итогового содержание отбирается по ***двум критериям***:

- ***полноты***: хороший тест должен проверять не менее 70% ДЕ, освоение которых предусмотрено в ГОС (рабочей программе);
- ***значимости***: в первую очередь должны проверяться ДЕ значимые с точки зрения содержания темы, раздела, дисциплины.

Значимость может быть установлена:

- по *частоте* встречаемости понятия, правила;
- *по времени*, которое отводит учебный план на изучение понятия, подраздела;
- на основании *экспертной оценки.*

Общее правило отбора содержания: чем выше значимость ДЕ, тем больше число заданий на ее проверку должен содержать тест.

Этап 4. *Разработка спецификации теста*

Проверяемые ДЕ размещаются в ***порядке убывания значимости*** и для каждой устанавливается количество ТЗ в соответствии с правилом этапа 3.

Результат удобно представить в виде таблицы.

Пример:

Пусть: t = 80 мин. ⇒ N = 40; пусть число ДЕ M = 4; ТЗ имеют 3 уровня трудности ($\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$ и $\boldsymbol{\alpha_3}$).

Таблица 1. – Доли и количество ТЗ на проверку каждой выделенной ДЕ

№ ДЕ	Доля от всех ТЗ	Кол-во ТЗ
ДЕ 1	*40%*	*16*
ДЕ 2	*30%*	*12*
ДЕ 3	*20%*	*8*
ДЕ 4	*10%*	*4*
	100%	***40***

Правило (рекомендация):

Количество ТЗ с НТ α_1 и α_2 – 60-70%, α_3 – 30-40%.

Для контроля уровней выше текущего максимум должен приходиться на ТЗ с α_2.

Таблица 2. – Количество ТЗ для каждой ДЕ в зависимости от трудности (спецификация теста)

Начальный вариант Табл. 2

ДЕ / α	ДЕ 1	ДЕ 2	ДЕ 3	ДЕ 4	Всего	Оценка доли
α_1	4	3	2	1	10	**30 (75%)**
α_2	8	4	5	3	20	
α_3	4	5	1	-	10	**10**
	16	**12**	**8**	**4**	**40**	**40**

Таблица 2. – Количество ТЗ для каждой ДЕ в зависимости от трудности (спецификация теста)

Конечный вариант Табл. 2

ДЕ / **α**	*ДЕ 1*	*ДЕ 2*	*ДЕ 3*	*ДЕ 4*	*Всего*	Оценка доли
α_1	*3*	*2*	*2*	*1*	*8*	***27 (67,5%)***
α_2	*7*	*5*	*4*	*3*	*19*	
α_3	*6*	*5*	*2*	*-*	*13*	***13***
	16	***12***	***8***	***4***	***40***	***40***

Тест содержит 12 (11) **типов** ТЗ

Иной вариант представления спецификации теста

ДЕ \ α	α_1	α_2	α_3	Всего
ДЕ 1	3	7	6	**16**
ДЕ 2	2	5	5	**12**
ДЕ 3	2	4	2	**8**
ДЕ 4	1	3	-	**4**
	8	**19**	**13**	**40**

Этап 5. *Подбор ТЗ с заданными дидактическими характеристиками*

Для каждого из выделенных на **этапе 4** типов ТЗ формулируется ***типовое задание*** с заданными характеристиками (*содержание, требуемый уровень усвоения, тип ТЗ* и пр.).

Дидактические характеристики типового задания:

- содержание (проверяемая ДЕ);
- уровень трудности;
- особенности представления задания:
  - ✔ тип тестового задания;
  - ✔ количество дистракторов;
  - ✔ характер концептов;
  - ✔ элементы оформления;
  - ✔ наличие графических объектов…

… и пр.

Этап 6. *Формирование фасетов*

Обеспечение вариативности теста при строгом сохранении его структуры и содержания – за счет **фасетного подхода**.

Фасет – группа ТЗ, с полностью идентичными дидактическими характеристиками (*та же ДЕ, тестовый тип, количество дистракторов, характер концептов, оформление*), полученная незначительным варьирование типового ТЗ..

Рекомендации:

- Для компьютерного тестирования группы обучаемых 20-30 чел. фасет должен содержать около **5** заданий, для тестирования курса (100 чел.) – **8-10** заданий.
- Задания фасета получаются **незначительным** варьированием содержания исходного (типового) ТЗ **при неизменной НТ и структуре задания** (типе, количестве дистракторов, характеру концептов, оформлению).
- При разработке компьютерного теста задания одного фасета должны сохраняться **в одной папке** (т.о., общее число папок оказывается равным N – числу заданий в тесте).

Этап 7. *Составление полной спецификации теста* (Таблица 3)

№ ТЗ	ДЕ	α	Тип ТЗ	t (мин)
1	*1*	*1*	*ОВ*	*1*
2	*1*	*1*	*ОВ*	*1*
3	*1*	*2*	*ОВ*	*2*
4	*2*	*1*	*О*	*1*
…	*…*	*…*	*…*	*…*
39	*4*	*3*	*С*	*3*
40	*4*	*3*	*П*	*3*
				80

Обозначения типов ТЗ:

- ОВ – одиночный выбор;
- МВ – множественный выбор;
- О – открытого типа;
- С – соответствия;
- П – установления последовательности

Этап 8. *Определение порядка интерпретации и использования результатов тестирования*

При разработке теста разработчиком должны быть ***установлены*** и ***задокументированы*** следующие позиции:

1) **Порядок оценки выполнения ТЗ**:
   - дихотомическая шкала (0-1, выполнено – не выполнено);
   - при использовании ТЗ типов МВ, С, П – доля правильности выполнения (от 0 до 1).
2) **Порядок оценки результата** тестирования (сырые баллы):
   - сумма набранных баллов по всем ТЗ;
   - сумма баллов по всем ТЗ с учетом весового множителя (НТ).

Этап 8. *Определение порядка интерпретации и использования результатов тестирования*

3) **Оценочная шкала** – порядок перевода сырых (первичных) баллов к принятой балльной шкале (5-ти, 100).
4) **Критерии результативности** – соответствие критериальных показателей интервалам на балльной шкале (***Примеры****: По Беспалько при ответе ≥ 70 % – полное усвоение; ГИБДД ≥ 90%; школа ≈ 40-50%; вступительные экзамены ≥ 30 % для участия в конкурсе*).
5) **Указания:** *для кого* предназначены результаты тестирования (уровень управления), *в течение какого времени* могут использоваться результаты (например, для сопоставления с подобными).

Этап 9. *Оценка качества и коррекция теста*

Производится по завершении тестирования с целью **совершенствования теста** как инструмента измерения.

Необходимые условия:

- достаточно большая статистика применения теста (*не менее 30 опрошенных*);
- представление результатов в табличном виде.

Характеристики:

- дискриминативность;
- надежность.

Характеристики качества теста: **дискриминативность**

Дискриминативность (*discriminatory power*) – дифференцирующая (различающая) способность тестовых заданий – характеристика, определяющаяся возможностью дифференцировать тестируемых с высокой и низкой результативностью выполнения теста.

Оценка дискриминативности (метод крайних групп):

1. Найти сумму баллов, набранных в тесте каждым протестированным; таблицу результатов переупорядочить в порядке убывания результатов.
2. Выделить «крайние» группы – отбирается от 10 до 30% лучших (N_u) и худших (N_d) по результатам выполнения всего теста.
3. Вычислить *индекс дискриминации* каждого задания по следующему правилу: пусть с заданием номер j справилось N_{uj} отвечавших из верхней группы и N_{dj} из нижней; тогда мера его дискриминативности D_j:

$$D_j = \frac{N_{uj}}{N_u} - \frac{N_{dj}}{N_d}$$

Пример:

Критерий: $D_j > 0,3$

Основные недостатки задач, оказавшихся непригодными:

- Излишняя сложность, запутанность формулировки;
- Неоднозначность условия;
- Очевидность решения;
- Абсурдность, нереальность вариантов ответов;
- Наличие нескольких правильных вариантов ответов, не оговоренных в условии;
- Зависимость результата от памяти или других индивидуальных особенностей испытуемых, а не от уровня развития умений и навыков, для оценки которых разрабатывалось задание.

Определение надежности

Тест является надежным, если он дает одни и те же результаты для каждого испытуемого при повторном тестировании (ретестовая надежность). Надежность теста характеризует точность теста как измерительного инструмента, устойчивость его к действию помех.

Методы оценки надежности:

1) повторное тестирование (подобные задания через некоторый промежуток времени – от 2-х недель до 6-ти месяцев);
2) параллельное тестирование (параллельной формой теста – деление испытуемых на 2 группы случайным образом, не более чем через неделю второе тестирование со сменой вариантов), расчет коэффициента корреляции;
3) расщепление (метод деления теста на 2 равноценные части по уровню сложности, трудности, на основе одинаковых коэффициентов дискриминации)

Характеристики качества теста: **надежность**

Надежность (*reliability*) – характеристика теста, отражающая точность педагогических измерений, степень постоянства результатов тестирования, а также устойчивость результа-тов теста к действию посторонних случайных факторов.

Надежность обусловливается:

- **помехоустойчивостью** теста к воздействию случайных факторов (*условия сдачи теста, личность тестирующего, психологическое состояние испытуемого и т. п.*);
- **согласованностью** отдельных заданий теста в целом.

Оценка надежности теста

(метод расщепленных частей теста):

1. Таблицу результатов разделить на две равные части (обычно по заданиям с четными и нечетными номерами).
2. Найти суммы баллов для каждого протестированного в каждой из частей.
3. Вычислить коэффициент линейной корреляции (Пирсона) между столбцами сумм (r_{1-2}).
4. Вычислить коэффициент надежности (коэффициент Спирмена-Брауна) ***H***:

$$H = \frac{2r_{1-2}}{1 + r_{1-2}}$$

Пример:

Градации надежности:

Неудовлетворительная **H < *0,7***

Удовлетворительная ***0,7 ≤ H < 0,8***

Хорошая ***0,8 ≤ H < 0,9***

Очень хорошая ***0,9 ≤ H < 0,95***

Отличная ***H ≥ 0,95***

Определение трудности

Индекс трудности:

$$U = (1 - \frac{n}{N}) \cdot 100$$

где *n* – число испытуемых, правильно решивших задачу, *N* – общее число испытуемых.

Для альтернативных заданий (с большой вероятностью угадывания ответа):

$$U = (1 - \frac{n - \frac{N_n}{m-1}}{N}) \cdot 100$$

где m – число вариантов ответа, N_n – число испытуемых, не решивших задачу.

Задания с 80 < U < 20 в тест не включаются – низкоинформативные задания (Гайда В.К., Захаров В. П.)

Определение валидности

Общее понимание – насколько полученные результаты соответствуют реальной действительности.

Очевидная валидность – с точки зрения испытуемого (доверительная валидность).

Содержательная валидность – соответствие содержания задания тому, что мы хотим проверить.

Критериальная валидность – способность теста служить индикатором или предсказателем определенной психической особенности, формы поведения человека; признак обоснованности, правомерности применения теста как измерительного инструмента.

Генеральный вывод:

Без правильного организованного жизненного цикла теста (построение спецификации, разработка типовых заданий, фасетирование, оценка качества материалов) **невозможна корректная интерпретация результатов тестирования !!!**

Благодарю за внимание!