

Ш.Б. Алтыбаева^{1*}, А.Н. Қуанышбекова², Ж. Абдулла³, А.Н. Есимбекова⁴

^{1,2,4} РГП на ПХВ «Национальный центр тестирования» МНВО РК, г. Астана, Республика Казахстан

³ КГУ «Специализированный лицей № 82 «Дарын»», г. Астана, Республика Казахстан

*e-mail: shygyla.altymbayeva@gmail.com

¹ORCID 0000-0003-0306-861X, ²ORCID 0009-0000-3780-6866,

³ORCID 0009-0004-9229-6108, ⁴ORCID: 0009-0005-3896-5654

ОЦЕНКА КОГНИТИВНОЙ СЛОЖНОСТИ ЗАДАНИЙ ПО ХИМИИ: ДАННЫЕ ЕНТ КАЗАХСТАНА

В настоящей работе оценивается, насколько две формы ЕНТ по химии сопоставимы не только по содержанию, но и по характеру познавательных действий, которые требуются от абитуриента. Проанализировано 80 заданий двух форм и классифицировали их когнитивную сложность по трехуровневой версии Depth of Knowledge (DOK) Уэбба. Шесть экспертов независимо кодировали все задания. На первом этапе абсолютное совпадение экспертных оценок составило 87%; случаи расхождения мы затем разбирали совместно и определяли итоговую категорию по консенсусу. В форме «Химия-1» к DOK-1, DOK-2 и DOK-3 мы отнесли соответственно 35,0%, 40,0% и 25,0% заданий, а в форме «Химия-2» - 57,5%, 25,0% и 17,5%. Наибольшее расхождение пришлось на DOK-1 и составило 22,5 процентного пункта. Дополнительно мы сопоставили DOK с эмпирической трудностью по модели Раша. Средняя трудность возрастала от DOK-1 к DOK-3, однако диапазоны отдельных уровней перекрывались. На основании этого результаты показывают, что одинаковое количественное распределение заданий по содержательным разделам еще не означает одинаковых когнитивных требований. В рамках данного пилотного анализа DOK-профиль оказался полезной дополнительной характеристикой тестовой формы. Поскольку мы рассмотрели только два варианта, выводы требуют проверки на более широком материале.

Ключевые слова: когнитивная сложность, DOK, ЕНТ, химия, high-stakes testing, валидность, модель Раша, параллельность форм.

Введение

В статье исходным положением является то, что для экзамена с высокими ставками недостаточно подтвердить надежность и соответствие теста содержательной спецификации. Важно также определить, какие действия должен выполнить абитуриент, чтобы получить правильный ответ. Этот аспект связан с представлением измеряемого конструкта и, следовательно, с обоснованностью интерпретации результатов [1, 2]. Если форма в основном проверяет воспроизведение фактов, она дает меньше информации о способности применять знания и рассуждать.

В исследовании разграничиваются когнитивная сложность и эмпирическую трудность задания. Под когнитивной сложностью под когнитивной сложностью понимается характер требуемого действия, тогда как трудность проявляется в результатах конкретной выборки. На р-value могут влиять формулировка, знакомость контекста, формат задания, подготовка участников и другие факторы [3, 4]. Поэтому когнитивный анализ когнитивный анализ рассматривается как самостоятельный слой описания теста. Сходную задачу решают исследования, в которых используются таксономия Блума и когнитивно-диагностические модели [5].

Для описания этого слоя мы выбрали модель Depth of Knowledge Уэбба и использовали трехуровневую схему: DOK-1 - воспроизведение, DOK-2 - применение знаний, навыков и понятий, DOK-3 - стратегическое мышление [6, 7]. При классификации учитывался не внешний вид задания и не число шагов решения, а то действие, без которого нельзя получить правильный ответ.

Возможность такого подхода к химическим заданиям показана, в частности, в работах Holme et al. и Raker et al. [8, 9].

В ЕНТ результаты тестирования используются при отборе абитуриентов в организации высшего образования. Поэтому, рассматривая параллельные формы, важно контролировать не только одинаковый охват тем. Даже при формальном соответствии одной спецификации формы могут различаться по тому, как часто они требуют воспроизведения, применения известного способа или самостоятельного выбора стратегии.

На материале химии это различие особенно удобно наблюдать. Часть заданий решается через воспроизведение определения, формулы или правила. В других случаях абитуриенту нужно сопоставить вещества и процессы, прочитать числовые данные, выбрать расчетный путь или объяснить химическое явление. Поэтому в анализе тематическая матрица не заменяет оценку характера познавательной работы.

При формировании аналитической схемы мы опирались на действующую спецификацию ЕНТ по химии. Она включает задания по строению и частицам вещества, закономерностям протекания реакций, энергетике, разделам «Химия и жизнь» и «Химия вокруг нас», а также расчетные и практико-ориентированные задания [10]. Это позволило нам отдельно рассматривать тематическое представление материала и требуемые от тестируемого действия.

При обзоре казахстанских публикаций мы обнаружили, что близкие по смыслу вопросы чаще рассматриваются в контексте учебной практики. Садыков, Кокибасова и Нурмыш, например, разработали разноуровневые задания для 9 класса, ориентированные на естественно-научную грамотность [11]. Авторы отмечают затруднения школьников при расчетах, составлении уравнений реакций и ответах на проблемные вопросы и предлагают различать задания по содержанию, характеру и объему работы.

Работа Шаңытбаевой также оказалась значимой для нашего обзора: в ней познавательные задания рассматриваются как средство формирования химических знаний [12]. Автор связывает их с развитием мышления, соединением теории и практики и элементами исследовательской деятельности. Для рассматриваемой задачи эта публикация важна прежде всего тем, что показывает интерес к характеру познавательного действия в казахстанском химическом образовании, хотя сравнение параллельных экзаменационных форм там не проводилось.

В международной литературе мы нашли примеры более формализованного анализа экзаменационных материалов. Okanlawon и Adeoti распределили 230 заданий пяти выпускных экзаменов West African Senior School Certificate Examination по категориям Revised Bloom's Taxonomy [13]. Такой дизайн позволил авторам количественно описать когнитивный состав реальных экзаменационных заданий, а не только учебных упражнений.

Еще один ориентир для нашей работы - исследование Zorluoğlu, Kizilaslan и Dönmez Yarıcıoğlu [14], где Revised Bloom's Taxonomy применена к программе и учебнику химии 9 класса. Авторы отнесли 61% результатов обучения к understand, 16% - к remember, 5% - к apply, 13% - к analyse, 0% - к evaluate и 5% - к create. Здесь важна сама возможность количественно описывать когнитивную структуру содержания.

Ren et al. анализировали пять форм китайского Гаокао по химии с помощью предметно-специфической модели компетенций [15] и выявили различия между формами по характеру требуемой деятельности. В качестве инструмента выбран - DOK, поскольку наша задача уже: сопоставить глубину требуемых действий в двух формах одного экзамена по единой шкале.

По результатам обзора мы выделили два основных направления: исследования учебных заданий по химии и работы, посвященные когнитивным требованиям реальных экзаменационных материалов. Для ЕНТ по химии нам не удалось обнаружить большого числа исследований, где несколько форм систематически кодировались бы по одной когнитивной шкале. Поэтому анализ сосредоточен на прямом сравнении двух реальных форм.

Практический смысл такого сравнения заключается. Спецификация показывает, какие темы и проектные уровни трудности должны быть представлены, но сама по себе не отвечает на вопрос, какую мыслительную операцию требует конкретное задание. С помощью DOK мы фиксировали эту характеристику отдельно от содержания и предполагаемой трудности.

Материалы и методы исследования

Исследование построено как пилотное сравнительное. Единицей анализа было отдельное задание ЕНТ по химии. Сначала мы кодировали задания двух форм по DOK, затем сопоставляли полученные когнитивные характеристики с содержанием, форматом и психометрическими показателями.

В анализ включены 80 заданий - по 40 из форм «Химия-1» и «Химия-2». По обеим формам были доступны экспертные оценки и данные выполнения заданий. Это позволило не ограничиваться содержательным описанием и проверить, как экспертно присвоенные уровни DOK соотносятся с наблюдаемой трудностью.

Использованы две реальные апробационные формы одного экзаменационного периода. Отбор был целенаправленным: включены варианты с полными эмпирическими данными, соответствовавшие действующей спецификации по содержанию и проектному распределению трудности. Случайную выборку форм мы на этом этапе не формировали.

Такое решение ограничивает область наших выводов. На основании двух форм результаты не переносятся на все варианты ЕНТ по химии; здесь мы прежде всего проверяем рабочую процедуру анализа. Каждая форма содержала задания с одним правильным ответом, контекстные задания, задания с несколькими ответами и задания на установление соответствия. Проектное соотношение категорий А, В и С составляло 50%, 30% и 20%.

В работе использовались сами тестовые задания, действующую спецификацию, контрольные листы разработки и внутренние материалы экспертной оценки Национального центра тестирования.

Операционализация когнитивной сложности

Для кодирования мы адаптировали модель Webb Depth of Knowledge [6, 7]. В рабочей версии оставили три уровня, которых, по нашему мнению, достаточно для различения основных типов действий в рассматриваемых заданиях по химии:

- DOK-1 - воспроизведение. К этому уровню мы относили задания, где правильный ответ можно получить за счет знания факта, определения, формулы, правила или выполнения простой освоенной процедуры.
- DOK-2 - применение. Этот уровень мы присваивали заданиям, в которых тестируемому нужно использовать знание или известный способ в конкретной либо измененной ситуации, интерпретировать, преобразовать или обработать представленную информацию.
- DOK-3 - стратегическое мышление. Сюда мы относили задания, где требуется связать несколько элементов, выбрать способ решения, выстроить содержательное рассуждение или применить стратегию, которая не задается напрямую стандартным алгоритмом.

При присвоении уровня учитывалось на минимально необходимое когнитивное действие, а не на формат задания. Категория А/В/С, объем текста и число расчетных шагов сами по себе уровень не повышали. Например, длинный расчет мы относили к DOK-2, если последовательность действий полностью задавалась известной процедурой.

Поэтому далее мы используем DOK как характеристику глубины действия, заложенного в задании, а не как другое обозначение эмпирической трудности

Экспертное кодирование

В экспертной процедуре участвовали шесть человек - специалисты Национального центра тестирования и обученные эксперты ETS. Каждый эксперт самостоятельно классифицировал все 80 заданий.

Перед началом работы мы предоставили экспертам единые описания трех уровней, адаптированные к химическому содержанию. В инструкции отдельно указали, что сложная формула, большой объем условия, несколько вычислений или проектная категория трудности сами по себе не являются основанием для присвоения более высокого DOK.

Мы предложили следующую последовательность кодирования: сначала эксперт формулировал действие, без которого задание нельзя решить, и только затем выбирал уровень DOK. Воспроизведение факта или правила относили к первому уровню; применение, интерпретацию и преобразование информации - ко второму; выбор стратегии и интеграцию нескольких взаимосвязанных элементов - к третьему.

Особенно внимательно мы рассматривали границу между DOK-2 и DOK-3. Мы не приравнивали многошаговое решение к стратегическому мышлению. Если учащийся мог выполнить задание по знакомой последовательности действий, мы оставляли его на уровне DOK-2.

По каждому заданию эксперт фиксировал выбранный уровень и краткое обоснование. Одновременно мы регистрировали содержательный раздел, способ измерения, проектную категорию A/B/C, тип стимульного материала и формат ответа. В дальнейшем эти признаки использовали для сопоставлений.

После независимого кодирования оценки были сведены в одну таблицу. Полное совпадение всех шести оценок на первом этапе наблюдалось в 87% случаев. Остальные задания проводилось совместное обсуждение: возвращались к формулировке задания и описанию требуемого действия, после чего принимали итоговую категорию по консенсусу.

Первоначальные оценки мы использовали для характеристики согласованности экспертной процедуры, а в последующий количественный анализ включали уже единый консенсусный код каждого задания.

Мы понимаем ограничение такой классификации: она описывает требование, заложенное в задании, но не фактический ход мысли каждого участника. Без когнитивных интервью нельзя утверждать, что все тестируемые решали одно и то же задание одинаковым способом. Поэтому DOK мы интерпретируем на уровне дизайна формы.

Анализ распределения когнитивной сложности. Для каждой формы мы подсчитали абсолютное число и процент заданий DOK-1, DOK-2 и DOK-3. Дополнительно объединили DOK-2 и DOK-3, чтобы получить простой описательный показатель того, насколько часто форма требует действий, выходящих за рамки непосредственного воспроизведения.

Разницу между формами мы выражали в процентных пунктах. Порог 10 процентных пунктов использовали только как внутренний ориентир для пилотного сравнения и не трактовали его как установленную границу эквивалентности форм.

Анализ проектной и эмпирической трудности. Проектную трудность определялась по спецификации в виде категорий А, В и С. Наблюдаемую трудность описывали через p -value - долю участников, которые дали правильный ответ.

Сначала мы сравнили распределения p -value внутри категорий А, В и С, а затем сопоставили те же категории с DOK. Принципиально различать эти признаки: А/В/С задает предполагаемую трудность на этапе конструирования, а DOK описывает характер требуемого познавательного действия.

Психометрический анализ. Для дополнительной характеристики исследуемых форм был проведен анализ на основе модели Раша [22] с использованием программного обеспечения Winsteps 5.8.3.0. Выборки абитуриентов, выполнявших формы «Химия-1» ($N=121$) и «Химия-2» ($N=161$), формировались независимо в рамках регулярного тестирования. Различие в объёме обусловлено организационными причинами и не связано с содержанием заданий. Для целей пилотного исследования такой разброс допустим. Размеры выборок различались, поэтому психометрические показатели двух форм интерпретировались преимущественно описательно [23, 24]. При данных объемах стандартные ошибки оценок трудности заданий различаются не более чем на 0,02 logit, что не влияет на интерпретацию

показателей соответствия (infit/outfit MNSQ) и не ставит под сомнение сопоставимость психометрических результатов двух форм.

Обеспечение качества и интерпретация результатов. Для контроля качества процедуры мы использовали три элемента: независимое кодирование шестью экспертами, единую инструкцию и обязательное обсуждение несовпадающих оценок. Наиболее внимательно мы контролировали границу DOK-2/DOK-3, поскольку именно здесь сложное содержание или длинный расчет могут привести к завышению уровня.

Психометрические показатели мы рассматриваем как вспомогательный источник информации. Они позволяют увидеть, как задания с разными DOK ведут себя в эмпирических данных, но не доказывают причинного влияния когнитивного уровня на трудность или надежность формы. По той же причине экспертный DOK не следует интерпретировать как прямое наблюдение реальных мыслительных процессов участников.

Результаты и обсуждение

В количественном анализе были использованы три уровня DOK: воспроизведение (DOK-1), применение знаний и навыков (DOK-2) и стратегическое мышление (DOK-3) [6, 7]. Краткие рабочие определения, которыми мы руководствовались, приведены в таблице 1.

Таблица 1 – Уровни DOK, использованные при анализе заданий по химии

Уровень	Когнитивная сложность
DOK 1	Вспомнить и воспроизвести
DOK 2	Применить знание, навык, понятие
DOK 3	Стратегическое мышление

Предварительно мы просмотрели распределение DOK по профильным предметам и увидели, что задания заметно различаются по характеру требуемых действий. В химии была представлена заметная доля DOK-3 (рисунок 1). Далее мы анализируем только две химические формы, поскольку именно по ним располагали полным набором экспертных кодов и эмпирических показателей. В обеих формах встречались задания от простого воспроизведения до задач, где требовалось анализировать связи и выбирать способ решения.

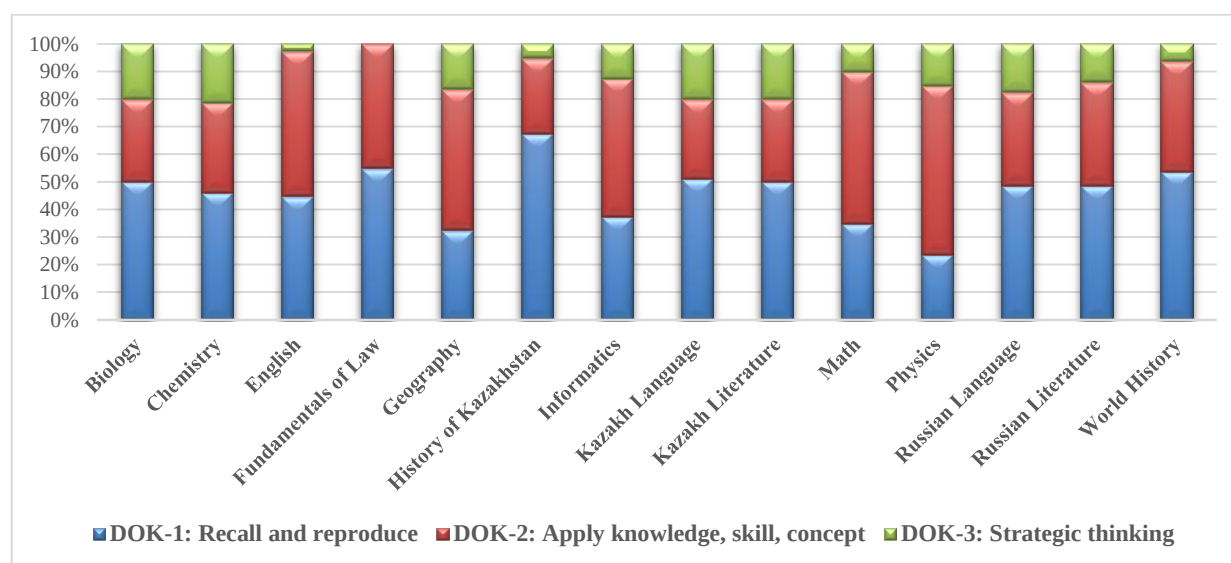


Рисунок 1 – Распределение заданий по уровням когнитивной сложности DOK

Такое сочетание уровней в химии мы считаем закономерным. Часть содержания действительно проверяется через факты, формулы и определения, но значительный блок заданий требует применять закономерности, читать числовую информацию, связывать вещества и процессы и выполнять расчеты. При этом наличие расчета само по себе мы не считали признаком DOK-3: решающим оставался характер рассуждения.

Задача сравнения разные предметы по единой «правильной» пропорции DOK. У химии, физики и биологии различаются предметные задачи и типичные способы решения. Поэтому интерес представляет не абстрактно высокий или низкий уровень когнитивной сложности, а согласованность профиля с целями конкретного теста.

Две формы сопоставлялись по итоговым экспертным кодам всех 40 заданий каждой формы.

Различие видно уже по частотам. В «Химии-1» мы отнесли к DOK-1 14 заданий (35,0%), к DOK-2 - 16(40,0%), к DOK-3 - 10 (25,0%). В «Химии-2» соответствующие значения составили 23 (57,5%), 10 (25,0%) и 7 (17,5%) (таблица 2).

Таблица 2 – Распределение уровней DOK по предмету «Химия» по данным содержательного анализа НЦТ

Профиль (№ формы)	DOK-1, %	DOK-2, %	DOK-3, %
Химия-1	35,0	40,0	25,0
Химия-2	57,5	25,0	17,5

Примечание: Cramér's $V = 0,23$ указывает на наличие связи между формой теста и распределением заданий по уровням DOK; величина эффекта интерпретируется как небольшая/умеренная в зависимости от принятого контекста и размера таблицы сопряженности

Если объединить обе формы, то из 80 заданий 37 (46,25%) классифицированы как DOK-1, 26 (32,50%) - как DOK-2 и 17 (21,25%) - как DOK-3. Все три уровня представлены, однако задания на воспроизведение образуют самую крупную группу.

Наиболее заметное расхождение получено по DOK-1. Во второй форме доля таких заданий выше на 22,5 процентного пункта: 57,5% против 35,0%. Для пилотного сравнения это превышает рабочий ориентир в 10 пунктов. Значение Cramér's $V = 0,23$ указывает, что различие распределений нельзя считать пренебрежимо малым; его необходимо проверить на большем числе форм.

Если объединить DOK-2 и DOK-3, получаем 65,0% в «Химии-1» и 42,5% в «Химии-2». Иными словами, первая форма чаще требует применения знаний или более сложного рассуждения, тогда как во второй заметно больше заданий на воспроизведение.

Важно, что это расхождение возникает при одинаковой тематической структуре. Содержательная матрица и когнитивный профиль рассматриваются и когнитивный профиль как разные стороны сопоставимости. При этом одного DOK также недостаточно, чтобы судить об эквивалентности форм: его нужно рассматривать вместе с содержательными и психометрическими данными.

Содержательный охват мы проверяли отдельно - по числу заданий в каждом основном разделе. Такой контроль показывает, выдержана ли тематическая матрица, но не раскрывает, какие действия выполняет тестируемый внутри этих тем. Такая сопоставимость предполагает согласованность содержательного представления, когнитивных требований, форматов заданий и других характеристик, определенных спецификацией теста.

Таблица 3 – Распределение заданий по содержательным разделам в двух формах теста по химии

Раздел	Номера заданий на форме 1	Число заданий	Номера заданий на форме 2	Число заданий
Частицы вещества	1, 2, 36	3	1, 2, 36	3
Закономерности протекания химических реакций	3, 4, 17	3	3, 4, 17	3
Энергетика химических реакций	5, 37	2	5, 37	2
Химия и жизнь	6-12, 31-33	10	6-12, 31-33	10
Химия вокруг нас	13-16, 34-35, 38-39	8	13-16, 34-35, 38-39	8
Задачи	18-25, 40	9	18-25, 40	9
Контекстные задания	26-30	5	26-30	5

Для параллельных форм ожидается сопоставимое представление содержательных областей, заданных спецификацией. Поэтому первым шагом сравнили фактическое число заданий по разделам.

Установлено, что в обеих формах число заданий совпадает по всем основным разделам (таблица 3). То есть по количественному содержательному охвату «Химия-1» и «Химия-2» различий не показали.

Однако после ДОК-кодирования выявлено, что совпадения по темам недостаточно для совпадения когнитивных профилей.

Это сопоставление стало одним из основных результатов: две формы могут содержать одинаковое число заданий по каждой содержательной области и при этом по-разному распределять воспроизведение, применение и стратегическое мышление.

Поэтому при сборке параллельных вариантов целесообразно проверять ДОК-профиль отдельно, не подменяя им содержательную экспертизу.

На следующем этапе мы сопоставили категории А, В и С с фактической трудностью и уровнями ДОК. В этот анализ вошли все задания обеих форм.

Спецификация задает соотношение 50/30/20 для категорий А, В и С. Для формы из 40 заданий это соответствует 20, 12 и 8 заданиям; именно это соотношение мы использовали как проектную основу.

Чтобы посмотреть, как проектные категории проявляются в данных, использовался показатель p-value. Распределения доли правильных ответов представлены на рисунке 2.

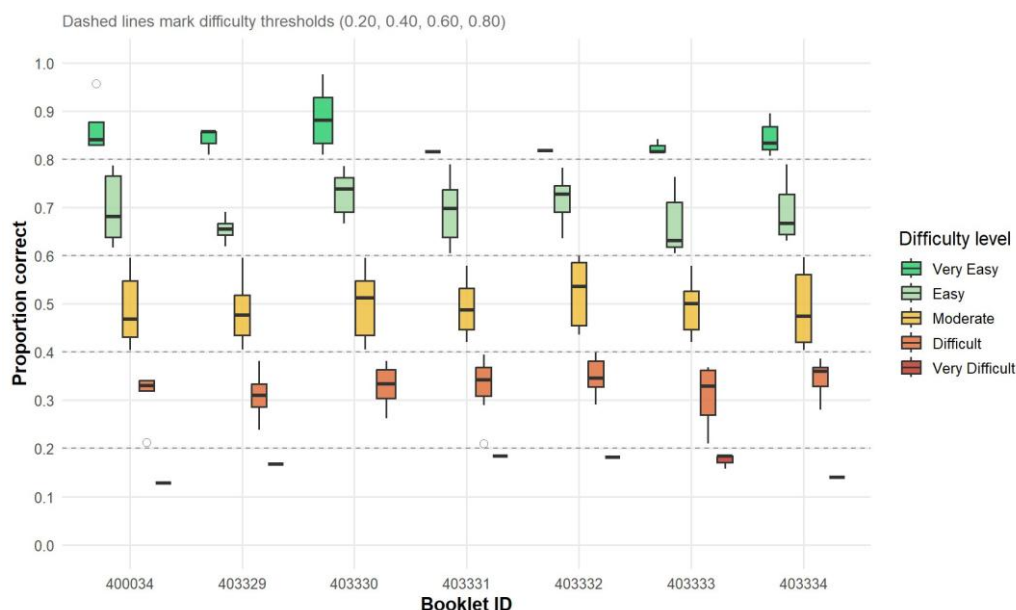


Рисунок 2 – Распределение p-значений заданий (доли правильных ответов) по проектным категориям трудности

Внутри каждой проектной категории мы обнаружили заметный разброс p-value. Распределения соседних категорий перекрываются, особенно А с В и В с С (рисунок 2). Поэтому не представляется возможным связать буквенную категорию с единственным диапазоном наблюдаемой трудности.

Одинаковая пропорция 50/30/20 в двух формах не приводит к одинаковому распределению p value внутри категорий. Следовательно, формальное совпадение проектной структуры само по себе не гарантирует совпадения эмпирической трудности.

При сопоставлении с DOK получена та же картина: проектная категория и когнитивный уровень описывают разные свойства задания. А/В/С фиксирует ожидаемую трудность на этапе разработки, а DOK - глубину действия, необходимого для решения. Поэтому более высокая категория трудности не интерпретируется как автоматический признак более высокого DOK.

При оценке параллельности эти показатели следует рассматривать отдельно: проектное распределение А/В/С, фактическую трудность и DOK-профиль.

Использование стимульного материала. Отдельно рассмотрен тип стимульного материала. В «Химии-1» текстовый или числовой стимул использовался в 60% заданий, в «Химии-2» - в 51%. Без отдельного стимула было 37% и 45% заданий соответственно, графика встречалась редко - в 3% и 4%.

Таблица 4 – Распределение типов стимульного материала в формах теста по химии

Профиль (№ формы)	Есть стимул — графический (% заданий)	Есть стимул — текстовый/числовой (% заданий)	Нет стимула (% заданий)
Химия-1	3	60	37
Химия-2	4	51	45

Если объединить обе формы, наблюдается следующая картина: 55,5% заданий содержат текстовый или числовой стимул, 41,0% не имеют отдельного стимульного материала и 3,5% используют графическое представление.

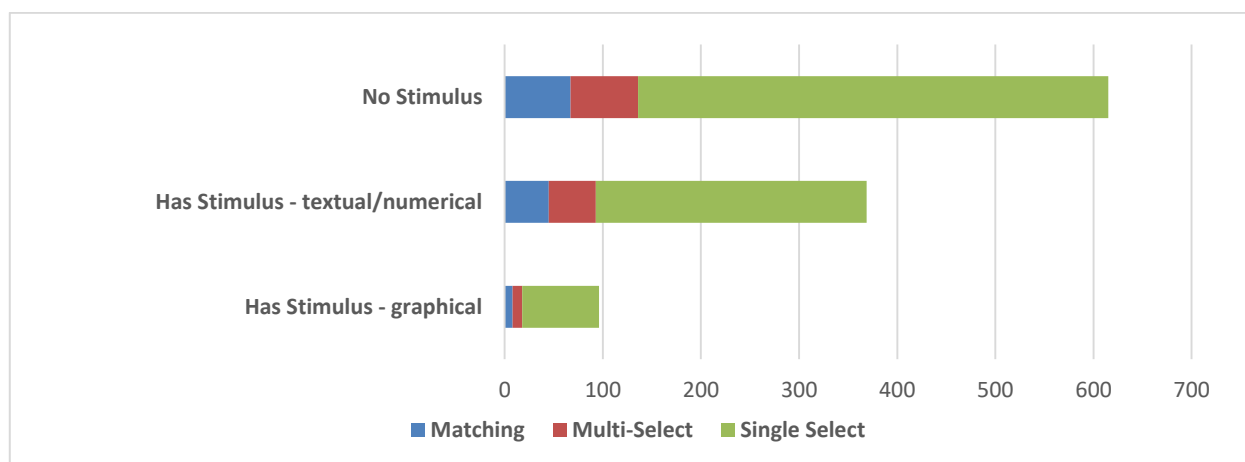


Рисунок 3 – Типы стимульного материала, включенные в задания профильного теста по химии

Мы не связываем преобладание текста и чисел с определенным уровнем DOK. Один и тот же тип стимула может сопровождать воспроизведение, применение известного способа или стратегическое рассуждение. При кодировании для нас было важно не наличие текста, числа или рисунка как такового, а то, что именно участник должен сделать с этой информацией.

Сопоставление психометрических характеристик двух форм

Психометрический анализ проводился отдельно для каждой формы в Winsteps 5.8.3.0. Для «Химии-1» использовали ответы 121 абитуриента, для «Химии-2» - 161. Эти показатели использовались не для прямого сравнения способностей двух выборок, а для описания поведения заданий и их сопоставления с экспертными DOK-кодами.

Таблица 5 – Сводные психометрические показатели форм «Химия-1» и «Химия-2»

Показатель	Химия-1 (N=121)	Химия-2 (N=161)
Надежность заданий (Item Reliability)	0.92	0.89
Надежность испытуемых (Person Reliability)	0.85	0.81
Cronbach's α	0.87	0.83
Разделимость заданий (Item Separation)	3.35	2.82
Средняя трудность заданий (logit)	0.00	0.00
Стандартное отклонение трудности заданий	0.71	0.68

Примечание. Средняя трудность заданий установлена равной 0.00 logit в результате центрирования, принятого в модели Раша. Стандартное отклонение отражает разброс трудностей относительно этого среднего значения.

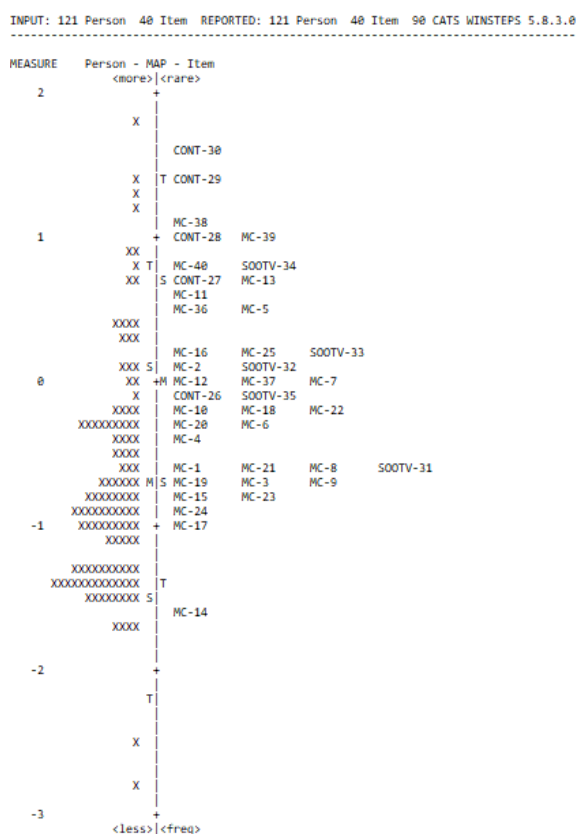
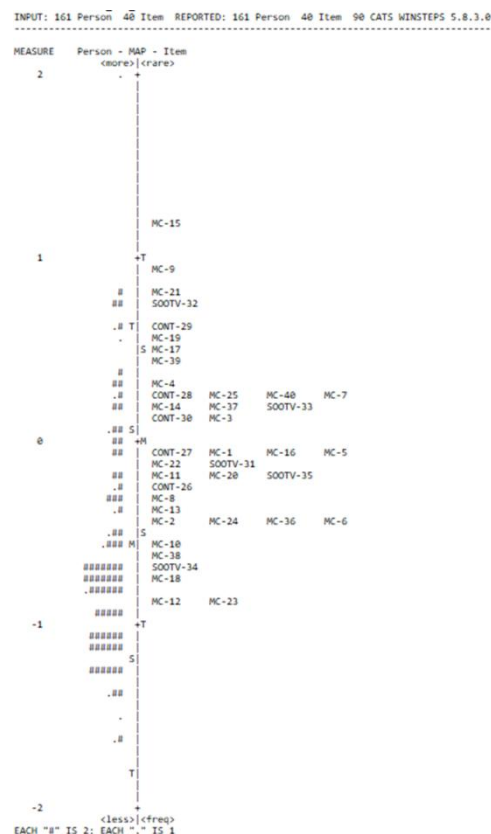
По заданиям обе формы показали высокую надежность: item reliability составила 0,92 и 0,89. Person reliability и item separation несколько выше для «Химии-1». Разброс трудностей почти одинаков - 0,71 против 0,68 logit. Поскольку формы выполняли разные группы, эти различия напрямую с DOK-профилем не связываются. Для более детального сопоставления задания каждой формы сгруппированы по итоговому DOK и рассчитаны средние значения MEASURE. Результаты приведены в таблице 6.

Таблица 6 – Средняя эмпирическая трудность (logit) заданий по уровням DOK

Уровень DOK	Химия-1	Химия-2
DOK-1	-0.37 (n=14)	-0.15 (n=23)
DOK-2	0.15 (n=16)	0.08 (n=10)
DOK-3	0.52 (n=10)	0.35 (n=7)

Примечание. Расчет произведен на основе колонки «MEASURE» из таблиц статистики заданий (TABLE 14.1 из Winsteps 5.8.3.0) для каждой формы. Группировка по уровням DOK выполнена согласно экспертной классификации.

В обеих формах наблюдается рост средней трудности от DOK-1 к DOK-3: разница между крайними уровнями составляет 0,89 logit для «Химии-1» и 0,50 logit для «Химии-2». Внутри групп сохраняется существенное перекрытие. Например, CONT-26, отнесенное к DOK-3, имеет трудность -0,14 logit, тогда как MC-30 с DOK-1 - 1,55 logit. DOK и эмпирическую трудность нельзя использовать как взаимозаменяемые показатели.

Рисунок 4 – Person - MAP - Item:
Химия-1Рисунок 5 – Person - MAP - Item:
Химия-2

Ту же картину мы видим на person-item maps. В «Химии-1» многие задания DOK-3 - CONT-29, CONT-30, MC-38, MC-39, MC-40, SOOTV-34 - располагаются примерно между 0,5 и 1,5 logit, а значительная часть DOK-1 - ниже. Но границы групп пересекаются: MC-30 (DOK-1; 1,55 logit) труднее CONT-26 (DOK-3; -0,14) и SOOTV-35 (DOK-3; -0,15). Во второй форме DOK-3 задания MC-15 (1,19) и MC-9 (0,97 logit) находятся в верхней части шкалы, однако положительные значения есть и у отдельных DOK-1 заданий, например MC-4 (0,31) и CONT-28 (0,28 logit).

По fit-статистикам большинство заданий оказалось в интервале 0,80-1,20 для infit и outfit MNSQ. В «Химии-1» внимание привлекли MC-16 (infit 1,34; ZSTD 3,28) и SOOTV-33 (1,37;

2,75). Во второй форме выделяются SOOTV-31 (1,20; 2,13) и MC-36 (1,18; 1,95). Эти задания — кандидаты для отдельного содержательного просмотра, прежде всего формулировок и возможных альтернативных стратегий решения. Единичные отклонения не меняют общей картины по формам.

В психометрических данных не обнаружено оснований считать DOK просто отражением трудности. Person reliability составила 0,85 и 0,81, item separation - 3,35 и 2,82, а разброс трудности почти одинаков. Две независимые выборки не позволяют утверждать, будто различия этих показателей вызваны именно составом DOK.

При проектировании параллельных форм целесообразно одновременно контролировать как минимум два аспекта: положение заданий на эмпирической шкале трудности и характер познавательных действий, необходимых для их выполнения.

Обсуждение

Главный результат исследования состоит в том, что две формы, совпадающие по числу заданий в содержательных разделах, не совпали по распределению DOK. Из этого следует ограниченный, но практический вывод: содержательная матрица описывает только одну сторону сопоставимости и не показывает, насколько похожа глубина работы с предметным материалом.

Наличие в обеих формах DOK-2 и DOK-3 рассматривается в контексте международного развития естественно-научного оценивания, где наряду с воспроизведением проверяются применение знаний, интерпретация данных, объяснение явлений и работа с моделями. Такие требования представлены в PISA, AP Chemistry и IB Chemistry [16, 17, 18]. В казахстанских исследованиях химического образования также обсуждается акцент на предметных компетенциях [19].

Второй результат - отсутствие прямого совпадения DOK с проектной категорией A/B/C и фактической трудностью. Перекрытие p-value между категориями и случаи, когда отдельное задание DOK-1 оказывается труднее DOK-3, показывают, что успешность решения зависит не только от глубины рассуждения. На результат могут влиять контекст, формулировка, объем информации и способ предъявления. Поэтому проектная трудность не является заменой когнитивному кодированию.

Анализ стимульного материала приводит к сходному выводу. Текст, числовые данные или графика сами по себе не повышают DOK; важен способ работы с ними. Для химии это особенно существенно, поскольку решение может требовать перехода между наблюдаемыми явлениями, субмикроскопическими представлениями и символическим языком [20, 21].

С практической точки зрения не предлагается просто увеличивать долю DOK-3. Более продуктивным является заранее задавать требуемое сочетание воспроизведения, применения и стратегического рассуждения и контролировать его вместе с тематической матрицей и трудностью. В таком случае DOK может стать дополнительным параметром спецификации, а не конкурирующей системой оценки качества.

Заключение

В нашем анализе две формы совпадали по количеству заданий в содержательных разделах, но различались по когнитивному профилю. Cramér's V для связи формы с уровнем DOK составил 0,23. Наиболее крупное различие мы получили по DOK-1: 57,5% во второй форме и 35,0% в первой, то есть 22,5 процентного пункта. Для пилотного сравнения это выше нашего рабочего ориентира в 10 пунктов. При этом мы не используем ни этот ориентир, ни Cramér's V как самостоятельное основание объявлять формы непараллельными, поскольку исследование охватывает только два варианта.

Полученные данные показывают, что показывают ограниченность оценки формы только по содержательной матрице и проектному соотношению категорий трудности.

Дополнительный DOK профиль помогает увидеть, какие типы познавательных действий представлены в форме и в какой пропорции.

Следующим этапом планируется включить в анализ большее число форм ЕНТ по химии. Это позволит проверить устойчивость найденных различий и перейти от пилотного описания к обоснованию целевого когнитивного профиля. Отдельно мы рассматриваем возможность проведения когнитивных интервью с учащимися, поскольку именно такой метод позволит сопоставить экспертно присвоенный DOK с фактическими стратегиями решения.

Список литературы

1. Messick S. Validity // Educational measurement / Ed. by R.L. Linn. - 3rd ed. - New York: American Council on Education / Macmillan, 1989. - P. 13-103.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for educational and psychological testing. - Washington, DC: AERA, 2014. - 230 p.
3. Gorin J.S., Embretson S.E. Item difficulty modeling of paragraph comprehension items // Applied Psychological Measurement. - 2006. - Vol. 30, No. 5. - P. 394-411. <https://doi.org/10.1177/0146621606288554>
4. Shavelson R.J. On the measurement of competency // Empirical Research in Vocational Education and Training. - 2010. - Vol. 2, No. 1. - P. 41-63. <https://doi.org/10.1007/BF03546488>
5. Ravand H., Shahi R., Effatpanah F., Moghadamzadeh A. Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's taxonomy // Frontiers in Education. - 2025. - Vol. 10. - Art. 1498765. <https://doi.org/10.3389/feduc.2025.1644811>
6. Webb N.L. Criteria for alignment of expectations and assessments in mathematics and science education (Research Monograph No. 6). - Madison, WI: Wisconsin Center for Education Research, 1997. - 45 p.
7. Webb N.L. Alignment, depth of knowledge, and change // Presentation at the 50th Annual Meeting of the Florida Educational Research Association. - Miami, FL, 2005. - 12 p.
8. Holme T., Knaus K., Murphy K., Blecking A. A valid and reliable instrument for cognitive complexity rating assignment of chemistry exam items // Journal of Chemical Education. - 2011. - Vol. 88, No. 5. - P. 554-560. <https://doi.org/10.1021/ed900070y>
9. Raker J.R., Trate J.M., Holme T.A., Murphy K. Adaptation of an instrument for measuring the cognitive complexity of organic chemistry exam items // Journal of Chemical Education. - 2013. - Vol. 90, No. 10. - P. 1290-1295. <https://doi.org/10.1021/ed400373c>
10. Национальный центр тестирования. Спецификация теста по химии для Единого национального тестирования. - Астана: НЦТ МНВО РК, 2025. - 12 с.
11. Садыков Т.М., Кокибасова Г.Т., Нурмыш С.Д. Development of multi-level tasks for the development of natural science literacy of students in chemistry // Gumilyov Journal of Pedagogy. - 2025. - Vol. 150, No. 1. - P. 42-57. <https://doi.org/10.32523/3080-1710-2025-150-1-42-57>
12. Шаңытбаева Ж.Н. Формирование химических знаний у обучающихся на основе познавательных заданий по химии // Yessenov Science Journal. - 2026. - № 1 (54). - С. 149-156. <https://doi.org/10.56525/km8rza83>
13. Okanlawon A.E., Adeoti Y.F. Content analysis of West African senior school certificate chemistry examination questions according to cognitive complexity // IFE Psychologia. - 2014. - Vol. 22, No. 2. - P. 1-12.
14. Zorluoğlu S.L., Kızılaslan A., Dönmez Yapıcıoğlu M. The analysis of 9th grade chemistry curriculum and textbook according to revised Bloom's taxonomy // Cypriot Journal of Educational Sciences. - 2020. - Vol. 15, No. 1. - P. 9-20. <https://doi.org/10.18844/cjes.v15i1.3516>
15. Ren D., Hu T., Dong X., Yao H. Assessing chemistry disciplinary competence in high-stakes examination forms: Evidence from the 2023 Chinese Gaokao // Frontiers in Education. - 2026. - Vol. 11. - Art. 1884670. <https://doi.org/10.3389/feduc.2026.1884670>
16. OECD. PISA 2025 science framework. - Paris: OECD Publishing, 2025. - 180 p.
17. College Board. AP Chemistry course and exam description (Effective Fall 2024). - New York: College Board, 2024. - 240 p.

18. International Baccalaureate Organization. Diploma Programme chemistry guide (First assessment 2025). - Cardiff: IBO, 2023. - 112 p.
19. Alzhanova G., Altynbayeva L., Zhanbulatova G., Tleulessova A., Toltebayeva F. Exploring competency-based chemistry teaching: Trends, challenges and perspectives // Chemistry-Didactics-Ecology-Metrology. - 2025. - Vol. 30, No. 1-2. - P. 87-97. <https://doi.org/10.2478/cdem-2025-0008>
20. Johnstone A.H. Why is science difficult to learn? Things are seldom what they seem // Journal of Computer Assisted Learning. - 1991. - Vol. 7, No. 2. - P. 75-83. <https://doi.org/10.1111/j.1365-2729.1991.tb00230.x>
21. Multiple representations in chemical education / Ed. by J.K. Gilbert, D. Treagust. - Dordrecht: Springer, 2009. - 439 p.
22. Wright B.D., Stone M.H. Best test design. - Chicago: MESA Press, 1979. - 238 p.
23. Linacre J.M. Sample size and item calibration stability // Rasch Measurement Transactions. - 1994. - Vol. 7, No. 4. - P. 328-332.
24. Smith A.B., Rush R., Fallowfield L.J., Velikova G., Sharpe M. Rasch fit statistics and sample size considerations for polytomous data // BMC Medical Research Methodology. - 2008. - Vol. 8, No. 1. - Art. 33. <https://doi.org/10.1186/1471-2288-8-33>

Ш.Б. Алтыбаева, А.Н. Қуанышбекова, Ж. Абдулла, А.Н. Есимбекова

ҰЛТТЫҚ БІРЫҢҒАЙ ТЕСТІЛЕУДЕГІ ХИМИЯ ТАПСЫРМАЛАРЫНЫҢ КОГНИТИВТІК КҮРДЕЛІЛІГІН БАҒАЛАУ: ҚАЗАҚСТАНДАҒЫ ҰБТ ДЕРЕКТЕРІ

Ірі ауқымды қабылдау емтихандарында екі нұсқаның бірдей тақырыптарды қамтуы олардың танымдық талаптары да бірдей дегенді білдірмейді. Осы пилоттық зерттеуде біз Қазақстанның ҰБТ химия пәні бойынша екі нұсқасындағы 80 тапсырманы талдадық. Тапсырмалардың когнитивтік күрделілігін Уэббтің Depth of Knowledge (DOK) моделінің үш деңгейлі бейімделуі арқылы бағаладық. Алты сарапшы барлық тапсырманы жеке кодтады; бастапқы абсолюттік келісім 87% болды, ал келіспеген жағдайларды кейін ортақ талқылауда нақтыладық. «Химия-1» нұсқасында DOK-1, DOK-2 және DOK-3 үлестері 35,0%, 40,0% және 25,0% болса, «Химия-2» нұсқасында 57,5%, 25,0% және 17,5% болды. Ең үлкен айырманы DOK-1 бойынша байқадық - 22,5 пайыздық тармақ. Раш моделі бойынша орташа эмпирикалық қиындық DOK-1-ден DOK-3-ке қарай өсті, бірақ деңгейлер арасындағы қиындық диапазоны толық бөлінген жоқ. Осы нәтижелерге сүйене отырып, біз мазмұндық бөлімдер бойынша тапсырмалар санының сәйкес келуі когнитивтік талаптардың да сәйкес болуына кепілдік бермейді деп есептейміз. DOK-профильді параллель нұсқаларды құрастыру мен сараптауда қосымша сипаттама ретінде пайдалануға болады. Зерттеу екі нұсқамен шектелгендіктен, бұл қорытындыларды кеңірек материалда тексеру қажет.

Түйін сөздер: когнитивтік күрделілік, DOK, ҰБТ, химия, жоғары ставкалы тестілеу, валидтілік, Раш моделі, нысандардың параллельдігі.

S.B. Altybayeva, A.N. Kuanyshbekova, Z. Abdulla, A.N. Yessimbekova

ASSESSMENT OF THE COGNITIVE COMPLEXITY OF CHEMISTRY TASKS IN THE NATIONAL ENTRANCE EXAMINATION: EVIDENCE FROM KAZAKHSTAN'S UNT

Two entrance-test forms may cover the same content while placing different cognitive demands on examinees. In this pilot study, we examined 80 chemistry items from two forms of Kazakhstan's Unified National Testing (UNT). We classified item demands using a three-level adaptation of Webb's Depth of Knowledge (DOK) framework. Six experts coded all items independently. Initial absolute agreement was 87%, and we resolved disagreements through subsequent discussion. In Chemistry-1, we classified 35.0% of the items as DOK-1, 40.0% as DOK-2, and 25.0% as DOK-3; in Chemistry-2, the corresponding proportions were 57.5%, 25.0%, and 17.5%. The largest difference was observed for DOK-1 and reached 22.5 percentage points. Our Rasch analysis showed an increase in mean empirical difficulty from DOK-1 to DOK-3, although the difficulty ranges overlapped. We therefore conclude that matching the number of items

across content areas does not necessarily produce matching cognitive profiles. We view DOK as an additional descriptor of parallel forms alongside content and psychometric information. Because our analysis covered only two forms, these findings should be treated as preliminary.

Keywords: cognitive complexity, DOK, UNT, chemistry, high-stakes testing, validity, Rasch model, test form parallelism.

References

1. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education / Macmillan.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (p. 230). AERA.
3. Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30(5), 394–411. <https://doi.org/10.1177/0146621606288554>
4. Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training*, 2(1), 41–63. <https://doi.org/10.1007/BF03546488>
5. Ravand, H., Shahi, R., Effatpanah, F., & Moghadamzadeh, A. (2025). Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's taxonomy. *Frontiers in Education*, 10, Article 1498765. <https://doi.org/10.3389/educ.2025.1644811>
6. Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6, p. 45). Wisconsin Center for Education Research.
7. Webb, N. L. (2005, November). *Alignment, depth of knowledge, and change* (p. 12) [Paper presentation]. 50th Annual Meeting of the Florida Educational Research Association, Miami, FL, United States.
8. Holme, T., Knaus, K., Murphy, K., & Blecking, A. (2011). A valid and reliable instrument for cognitive complexity rating assignment of chemistry exam items. *Journal of Chemical Education*, 88(5), 554–560. <https://doi.org/10.1021/ed900070y>
9. Raker, J. R., Trate, J. M., Holme, T. A., & Murphy, K. (2013). Adaptation of an instrument for measuring the cognitive complexity of organic chemistry exam items. *Journal of Chemical Education*, 90(10), 1290–1295. <https://doi.org/10.1021/ed400373c>
10. Natsional'nyy tsentr testirovaniya. (2025). *Spetsifikatsiya testa po khimii dlya Yedinogo natsional'nogo testirovaniya* [Chemistry test specification for the Unified National Testing] (p. 12). NTsT MNVO RK.
11. Sadykov, T. M., Kokibasova, G. T., & Nurmysh, S. D. (2025). Development of multi-level tasks for the development of natural science literacy of students in chemistry. *Gumilyov Journal of Pedagogy*, 150(1), 42–57. <https://doi.org/10.32523/3080-1710-2025-150-1-42-57>
12. Shañytbaeva, Zh. N. (2026). Formirovanie khimicheskikh znaniy u obuchayushchikhsya na osnove poznavatel'nykh zadaniy po khimii [Formation of chemical knowledge in students based on cognitive chemistry tasks]. *Yessenov Science Journal*, 1(54), 149–156. <https://doi.org/10.56525/km8rza83>
13. Okanlawon, A. E., & Adeoti, Y. F. (2014). Content analysis of West African senior school certificate chemistry examination questions according to cognitive complexity. *IFE Psychologia*, 22(2), 1–12.
14. Zorluoğlu, S. L., Kızılaslan, A., & Dönmez Yapıcıoğlu, M. (2020). The analysis of 9th grade chemistry curriculum and textbook according to revised Bloom's taxonomy. *Cypriot Journal of Educational Sciences*, 15(1), 9–20. <https://doi.org/10.18844/cjes.v15i1.3516>
15. Ren, D., Hu, T., Dong, X., & Yao, H. (2026). Assessing chemistry disciplinary competence in high-stakes examination forms: Evidence from the 2023 Chinese Gaokao. *Frontiers in Education*, 11, Article 1884670. <https://doi.org/10.3389/educ.2026.1884670>
16. OECD. (2025). *PISA 2025 science framework* (p. 180). OECD Publishing.
17. College Board. (2024). *AP Chemistry course and exam description (Effective Fall 2024)* (p. 240). College Board.
18. International Baccalaureate Organization. (2023). *Diploma Programme chemistry guide (First assessment 2025)* (p. 112). IBO.

19. Alzhanova, G., Altynbayeva, L., Zhanbulatova, G., Tleulessova, A., & Toltebayeva, F. (2025). Exploring competency-based chemistry teaching: Trends, challenges and perspectives. *Chemistry-Didactics-Ecology-Metrology*, 30(1–2), 87–97. <https://doi.org/10.2478/cdem-2025-0008>
20. Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning*, 7(2), 75–83. <https://doi.org/10.1111/j.1365-2729.1991.tb00230.x>
21. Gilbert, J. K., & Treagust, D. (Eds.). (2009). *Multiple representations in chemical education* (p. 439). Springer.
22. Wright, B. D., & Stone, M. H. (1979). *Best test design* (p. 238). MESA Press.
23. Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328–332.
24. Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, 8(1), Article 33. <https://doi.org/10.1186/1471-2288-8-33>

Сведения об авторах:

Алтыбаева Шугыла Болатовна (автор-корреспондент) - РГП на ПХВ «Национальный центр тестирования» МНВО РК, руководитель управления, Астана, Казахстан, e-mail: shygyla.altybayeva@gmail.com

Куанышбекова Арайлым Ниязбекқызы - РГП на ПХВ «Национальный центр тестирования» МНВО РК, ведущий эксперт, Астана, Казахстан, e-mail: quanyshbekova98@mail.ru

Абдулла Жанар - КГУ «Специализированный лицей № 82 «Дарын»», учитель химии, Астана, Казахстан, e-mail: jeneiro@mail.ru

Есимбекова Аруна Нурлыбаевна - РГП на ПХВ «Национальный центр тестирования» МНВО РК, ведущий эксперт, Астана, Казахстан, e-mail: aruna.iesimbekova@mail.ru

Авторлар туралы мәлімет:

Алтыбаева Шугыла Болатовна (автор-корреспондент) - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, бөлім басшысы, Астана, Қазақстан, e-mail: shygyla.altybayeva@gmail.com

Куанышбекова Арайлым Ниязбекқызы - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, жетекші сарапшы, Астана, Қазақстан, e-mail: quanyshbekova98@mail.ru

Абдулла Жанар - «№ 82 «Дарын» мамандандырылған лицейі» КММ, Химия пәні мұғалімі, Астана, Қазақстан, e-mail: jeneiro@mail.ru

Есимбекова Аруна Нурлыбаевна - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, жетекші сарапшы, Астана, Қазақстан, e-mail: aruna.iesimbekova@mail.ru

Information about the Authors:

Altybayeva Shugyla (corresponding author) - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Head of Department, Astana, Kazakhstan, e-mail: shygyla.altybayeva@gmail.com

Kuanyshbekova Arailym - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Leading Expert, Astana, Kazakhstan, e-mail: quanyshbekova98@mail.ru

Abdulla Zhanar - KSU «Specialized Lyceum No. 82 «Daryn»», Chemistry Teacher, Astana, Kazakhstan, e-mail: jeneiro@mail.ru

Yessimbekova Aruna - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Leading Expert, Astana, Kazakhstan, e-mail: aruna.iesimbekova@mail.ru