

# Так ли полезна психометрика для академической психологии?

Юлия Тюменева

Статья поступила в редакцию в феврале 2023 г. Тюменева Юлия Алексеевна — кандидат психологических наук, старший научный сотрудник Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: jutu@yandex.ru. ORCID: <https://orcid.org/0000-0002-2381-917X>

**Аннотация** Психологические теории относительно способностей и личностных черт часто полагаются на результаты психометрического моделирования. Предполагается, что оно связывает ответы на задания теста с ненаблюдаемым «конструктом» (чертой, способностью), который и «моделируется» на основе данных теста. Однако свидетельствует ли согласие между данными и моделью о том, что модель репрезентирует психологический конструкт? Насколько вообще психометрическое моделирование является моделированием в общенаучном значении этого термина? От ответа на эти вопросы зависит обоснованность использования данных моделирования для понимания психологических феноменов. В статье анализируется логика психометрического моделирования в сравнении с моделированием в других науках и утверждается, что психологические феномены как предмет моделирования не участвуют ни в построении, ни в коррекции моделей. Автор поднимает проблему необоснованных интерпретаций результатов моделирования в психологии и их нежелательных последствий для психологической теории. При этом значение психометрического моделирования как инструмента для решения управленческих задач еще ждет своей оценки.

**Ключевые слова** психометрическое моделирование, моделирование латентного конструкта, психологический конструкт, психологическая теория, тест

**Для цитирования** Тюменева Ю.А. (2023) Так ли полезна психометрика для академической психологии? *Вопросы образования / Educational Studies Moscow*, № 3, сс. 197–220. <https://doi.org/10.17323/vo-2023-16781>

## Is Psychometrics So Useful for Academic Psychology?

Yulia Tyumeneva

Yulia A. Tyumeneva — PhD, Associate Professor and Senior Research Fellow in Institute of Education, National Research University, Higher School of Economics, Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: jutu@yandex.ru. ORCID: <https://orcid.org/0000-0002-2381-917X>

**Abstract** Psychological theories regarding ability and personality traits often rely on the results of psychometric modelling. The latter is assumed to link responses to test items to an unobserved 'construct' (trait, ability), which is 'modelled' from the test data. However, does the agreement between the data and the model indicate that the model represents a psychological construct? To what extent is 'psychometric modelling' modelling in the general scientific sense of the term? The validity of using modelling data to understand psychological phenomena depends on the answer to these questions. The article analyses the logic of psychometric modelling in comparison with modelling in other sciences and argues that psychological phenomena as a subject of modelling are not involved neither in the construction nor in the correction of models. The problem of unjustified interpretations of modelling results in psychology and their undesirable consequences for psychological theory is raised. At the same time, the use of psychometric modelling for human resource decision-making is still waiting for its evaluation.

**Keywords** psychometric modelling, latent construct modelling, psychological construct, psychological theory, test

**For citing** Tyumeneva Yu.A. (2023) Tak li polezna psikhometrika dlya akademicheskoy psikhologii? [Is Psychometrics So Useful for Academic Psychology?]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 197–220. <https://doi.org/10.17323/vo-2023-16902>

### **Предварительные замечания**

Предлагаемая вниманию читателей статья вызвала довольно острую полемику между мной и одним из моих рецензентов. При том что я не могу согласиться со многими его аргументами, я считаю состоявшееся обсуждение чрезвычайно важным, поскольку оно подсвечивает всю сложность затронутой проблемы и отсутствие ее однозначного решения. Чтобы проследить логику доводов обеих сторон, мы согласились на их полную публикацию в будущем номере журнала. Таким образом ничье авторство не нарушается, а читатель может получить удовольствие от интригующей дискуссии. Я благодарна обоим рецензентам и редакторам за публикацию этой статьи, несмотря на то что она очевидно диссонирует с их личной позицией, а также, возможно, с другими материалами этого выпуска.

### **Введение**

Психометрика, по определению Американской психометрической ассоциации, занимается количественной оценкой и измерением психических качеств, поведения, успешности и т.п., а также разработкой, анализом и совершенствованием тестов, опросников и других инструментов, используемых для такого измерения<sup>1</sup>. Более аккуратное определение предлагает Международная энциклопедия социальных и поведенческих наук: психометрика — это научная дисциплина, занимающаяся вопросами конструирования психометрических моделей психо-

<sup>1</sup> APA Dictionary of Psychology: <https://dictionary.apa.org/psychometrics>

логических данных [Borsboom, Molenaar, 2015]. В этих моделях теоретический конструкт, например интеллект, систематически координируется с наблюдениями, например с баллами по тесту интеллекта. Чаще всего это делается с помощью так называемых моделей латентных переменных, которые работают как общие детерминанты набора баллов по тесту<sup>2</sup>.

Идея моделирования как метода изучения психологических конструктов пронизывает всю современную психометрическую литературу. Утверждения о том, что психометрика может моделировать латентный психологический конструкт, звучат в литературе совершенно явно. Выводы или гипотезы о конструкте делаются по результатам психометрического моделирования в отношении самых разных феноменов. Идет ли речь о мотивах [Freund, Lohbeck, 2021], настойчивости [Credé, 2018; Тюменева, Kardanova, Kuzmina, 2019], личности [Franić et al., 2014; Walton et al., 2008; Streckert, Kurtz, Kajonius, 2023], компетентности [Hartig, Höhler, 2009], мышлении [Dumas, Dong, 2022; Wagner, Harvey, 2006; Alexander et al., 2016; Zhao, Alexander, Sun, 2021; Araujo et al., 2019], креативности [Qian, Plucker, Yang, 2019; Shaw, Карнек, Morelli, 2021], субъективном благополучии [Nima et al., 2020], эмоциях [Lange et al., 2020; Power, 2006], установках [Hauwaert van, Schimpf, Azevedo, 2020] — во всех случаях психометрические модели составляют существенную часть методического инструментария. Заявления о возможности предоставлять информацию об измеряемом конструкте — мышлении, мотивах, установках и т.д., — более или менее явно сформулированные, можно найти во многих учебниках по психометрике и в методической периодике [Ackerman, Gierl, Walker, 2003; Fox, 2005; Linden van der, Hambleton, 2013; Nering, Ostini, 2010; Sijtsma, Ark van der, 2020]. Моделирование латентных конструктов стало дополнительным способом валидации шкал, в том числе использующихся в международных программах оценки качества образования, таких как PISA или PIRLS, средством поддержки интерпретации результатов тестирования в терминах черт и способностей и даже источником новых конструктов и теорий [Buchholz, Hartig, 2020; Kunina-Habenicht, Goldhammer, 2020].

У термина «конструкт» до сих пор нет исчерпывающего и общепринятого определения, при этом из-за смешивания с близкими понятиями, такими как «атрибут», «фактор», «ла-

<sup>2</sup> Не все психометрические модели имеют дело с латентными переменными, например есть сетевые или формативные модели, когнитивные диагностические модели [Templin, Henson, 2006; Schmittmann et al., 2013]. В этой работе мы будем обсуждать моделирование конструктов в парадигме современной теории тестирования и факторного анализа, хотя эта же логика применима и к другим моделям.

тентная переменная», «ненаблюдаемая переменная», его содержание становится еще более туманным. Не будет ошибкой утверждать, что часто под конструктом подразумевается предположительная ненаблюдаемая детерминанта какого-то наблюдаемого паттерна поведения, в том числе и выполнения заданий теста. Способности, черты характера, установки — все эти характеристики описываются как латентные конструкты. Латентные в том смысле, что они не наблюдаются непосредственно, но предположительно объясняют поведение.

**Латентный /  
психоло-  
гический /  
ненаблюдаемый  
конструкт**

Из-за отсутствия однозначного определения понятия «конструкт» его включение в профессиональный жаргон имело противоречивые последствия. С одной стороны, понятие «конструкт» позволило обойти ограничения операционализма, открыв доступ к теоретическим обобщениям в количественной психологии, а с другой — затруднило коммуникацию, смешав языки статистического вывода, лингвистических концепций и психологических теорий [Cronbach, Meehl, 1955; MacCorquodale, Meehl, 1948; Markus, Borsboom, 2013; Michell, 2013].

Перечислим типичные значения, в которых термин «конструкт» употребляется в психометрической литературе, в том числе в учебниках<sup>3</sup>. Во-первых, понятие «конструкт» используется, чтобы репрезентировать вещи (объекты, процессы, механизмы, состояния, отношения), объективно существующие, но принципиально ненаблюдаемые, независимо от способности ученого их заметить. Например, интеллект может мыслиться как реальность, влияющая на ответы по тесту способностей и доступная изучению через эти ответы. Во-вторых, термин «конструкт» может использоваться как чисто лингвистическое образование, необходимое для обозначения классов объектов, процессов или состояний, используемое в том числе для нужд познания или деятельности. В этом значении «интеллект» является понятием, обобщающим умения и знания, нужные для решения задач теста. Наконец, термин «конструкт» может использоваться в техническом смысле для описания и объяснения статистических феноменов, например когда нужно охарактеризовать распределение переменных. Здесь он часто подменяет понятия «фактор», «латентная переменная» и «размерность». Именно в этом значении Ч. Спирмен ввел понятие общего фактора интеллекта. В последнее время предпринимаются попытки определить конструкт как эмерджентное понятие [Lange et al., 2020]. В общем, такая непроясненность важнейше-

<sup>3</sup> Интереснейшее описание значений, в которых употребляется термин «конструкт», можно найти в специальном выпуске журнала *New Ideas in Psychology* (№ 1 за 2013 г.).

го психометрического термина имеет серьезные последствия для интерпретации результатов моделирования конструкторов в психометрике [Marau, Gabriel, 2013].

Имея в виду неопределенность термина «конструкт», будем все же считать, что в заявлениях о возможности построить «измерительную модель латентного конструкта» последний понимается как некоторый психологический феномен (черта, способность, установки и проч.), лежащий за пределами теста, детерминирующий ответы на тест и за счет этого доступный для моделирования и измерения. В этой статье «конструкт» будет взят именно в этом, весьма распространенном, значении. Например,

«в психологии и образовании, IRT — один из главных инструментов измерения способностей и установок» [Linden van der, Hambleton, 2013]; «целью является проиллюстрировать, как пользователи теста и исследователи могут применять многомерную IRT (MIRT) для того, чтобы понять, что их тесты измеряют...» [Ackerman, Gierl, Walker, 2003]; «обсуждаются выводы для исследования критического мышления на основе измерений в рамках IRT» [Wagner, Harvey, 2006]; «Является ли креативность специфической или общей способностью? В этом исследовании использованы многоуровневые эксплораторные IRT модели <...> Результаты предполагают, что креативность является скорее общей, а не специфической способностью» [Qian, Plucker, Yang, 2019].

Насколько обоснованны подобные утверждения? Эта статья посвящена анализу логики психометрического моделирования как способа репрезентировать латентный конструкт или обнаружить его структуру. Будет показано, что, во-первых, то, что в психометрике получило название «моделирование», или «построение измерительной модели», имеет весьма отдаленное отношение к моделированию как к общенаучному методу; во-вторых, что психометрическое моделирование в принципе не может дать информацию, которую можно было бы использовать для понимания того феномена, который подлежит моделированию.

### **Моделирование как метод исследования**

Поскольку моделирование — общенаучный метод, применяемый для решения широкого круга задач, например в экологии [Arhonditsis et al., 2006], эволюционной биологии [Pugesek, Tomer, von Eye, 2003], генетике [Franić et al., 2012], медицине [Ottensen, 2000], в исследованиях экономического поведения [Birnbbaum, 2008], спорте, рассмотрим общую логику этого метода.

Любая модель используется, чтобы описывать, объяснять и предсказывать поведение реальной системы. Во всех случаях моделируемый феномен тщательно изучается на предмет внутренних закономерностей его существования. Эти закономерности должны быть репрезентированы моделью, поэтому именно результаты предварительных исследований составляют параметры модели. Далее на основе сравнения расчетных данных с реальными моделью корректируется или выбирается лучший ее вариант. Параметры, которые не повышают точность прогноза, будут, скорее всего, исключаться из модели. Поэтому обратная связь от поведения реальной системы абсолютно необходима для моделирования и коррекции модели. Ясно также, что эта обратная связь относительно точности модели должна генерироваться независимо от модели [Oberkampff et al., 2022].

Проиллюстрируем принципы моделирования и его возможности двумя примерами. В климатологии математические модели эффективны для проверки предположений о факторах, приводящих к тем или иным экологическим последствиям. Например, *Soil Water Atmosphere Plant (SWAP)* моделирует перенос воды и тепла в зоне грунтовых вод во взаимодействии с развитием растительности [Айзель, Гусев, Насонова, 2017; Dam van et al., 2008]. В модели используется уравнение, которое включает показатели корневой экстракции воды, гидрофильности почвы (которая зависит, в свою очередь, от глубины ее промерзания), движения почвенной влаги, водосброса, а также теплоемкости и теплопроводности почвы. Для оценки переноса воды и растворенных веществ SWAP учитывает основные количественные закономерности процессов конвекции, дисперсии, адсорбции и др., также известные до и независимо от модели. В итоге модель позволяет объяснить явления засухи на изучаемой территории, а также обеспечение влагой растений и их рост. Модель многократно тестировалась в отношении разных климатических условий и показала свои преимущества перед другими моделями в прогнозе многих событий.

Другой пример: в фигурном катании, чтобы проверить гипотезы о влиянии специфических движений на высоту прыжка, может использоваться его математическая модель. Для этого сначала собирается информация о параметрах прыжка, например из видеозаписей его выполнения. У этих же фигуристов дополнительно оценивается сила плеч и мышц ног при разных угловых скоростях и типичная высота прыжков [Podolsky et al., 1990]. На основе корреляционных данных выявляются параметры, наиболее сильно связанные с высотой прыжка, и используются для параметризации математической модели. Качество модели оценивается по ее предсказательной силе в отношении

высоты реальных прыжков на соревнованиях, т.е. в отношении информации, полученной независимо от модели. Данные моделирования находят применение при разработке программ силовых тренировок для фигуристов, а также при прогнозировании риска ошибок или падений при выполнении прыжков [Rhodes, Putkaradze, 2022].

Психометрическое моделирование латентных конструктов иногда описывается и в другой логике — логике выявления ненаблюдаемых объектов [Maraun, 2017]. Она лежит в основе методологии изготовления и коррекции детекторов, самые известные из них — лакмусовая бумага и металлодетектор. В самом деле, психометрика прибегает к терминологии, близкой именно этой логике: «наблюдаемая переменная», «латентный конструкт», «манифестируемая переменная», «индикатор» и проч. Что пытается сделать психометрика по отношению к некоторой ненаблюдаемой реальности — это выработать и обосновать метод детекции этой реальности.

М. Мараун выделяет следующие принципы в технологии создания и работы любого детектора.

1. При необходимости принять решение относительно присутствия специфического ненаблюдаемого объекта создается протокол обнаружения.
2. Такой протокол включает спецификацию объектов класса  $U$  путем установления правила, определяющего свойства, которыми должен обладать объект, чтобы считаться элементом  $U$ .
3. Задаются логические суждения, которые связывают наличие элемента  $u$  из  $U$  с наблюдаемым «сигналом»  $O$ . Возможны три типа таких суждений:
  - «Если  $U$ , тогда  $O$  ( $O$  — необходимое условие  $u$ )»;
  - «Если  $O$ , тогда  $U$  ( $O$  достаточное условие  $u$ )»;
  - « $U$  присутствует, если и только если  $O$  ( $O$  — необходимое и достаточное условие  $u$ )».
4. Инструмент обнаружения (детектор) позволяет принять решение о наличии или отсутствии  $u$ . Детектор — это реализация конкретного логического суждения.
5. Определяются дополнительные условия, которые должны быть выполнены для того, чтобы детектор функционировал должным образом [Ibid.].

Проиллюстрируем эту логику на примере работы металлодетектора. В этом случае объектами обнаружения являются металлические предметы, т.е. элементы, обладающие свойствами металлов, а правила определения металлов известны заранее

и независимо от детектора. Суждение о присутствии искомого объекта относится к третьему типу: электромагнитный импульс определенной длительности возникает тогда и только тогда, когда вблизи есть металлический объект. Импульсно-индукционный металлодетектор является реализацией этого логического суждения. Побочные условия: импульсно-индукционный металлодетектор не работает вблизи телевизоров, радиоприемников, сотовых телефонов и других устройств, производящих радиоволны [Maraun, 2017].

Обобщая вышесказанное, можно заключить, что моделирование и/или детекция как общенаучные методы исследования требуют наличия предварительно установленных количественных закономерностей в эмпирической системе (объекте, процессах, взаимодействиях и проч.), в отношении которой строится модель или детектор. Они нужны для параметризации модели или установления правил принятия решения при обнаружении сигналов детектором, а также для понимания граничных условий работоспособности модели или детектора. Кроме того, обратная связь для оценки модели должна обеспечиваться независимо от данных, поставляемых этой моделью. Так, информация о наличии металла должна поступить независимо от данных детектора, чтобы можно было оценить эффективность последнего.

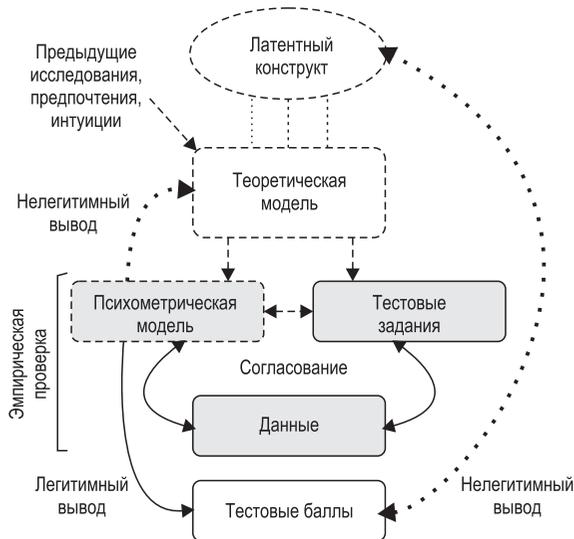
### **Моделирование в психометрике**

Современные психометрические методы моделирования очень разнообразны (обзор см. в [Sen, Cohen, 2019]). Для целей статьи мы остановимся на моделировании латентных конструктов в рамках современной теории тестирования (IRT-моделирование), близкородственной факторному анализу. Однако описываемая ниже логика во многом распространяется и на другие типы психометрических моделей, в том числе сетевые.

Процесс моделирования можно описать как последовательность шагов или содержательных блоков (рис. 1). Первый, предварительный, шаг — это просто название и описание психологического феномена (конструкта) или процесса, который должен быть замоделирован и измерен. К примеру, научное мышление может быть описано как мышление с тенденцией формулировать суждения относительно наблюдаемых явлений в форме альтернативных гипотез и условий принятия этих гипотез после проверки. Второй шаг — это формулирование психологической теории или гипотезы относительно природы конструкта, его аспектов, структуры и компонентов, внутренних связей, а также взаимодействия этого конструкта с другими психологическими феноменами или объективными

событиями в жизни человека. Такая теоретическая модель относительно природы «латентного конструкта» может строиться на основе предыдущих исследований, наблюдений, интуиций или суждений экспертов об этой черте/способности<sup>4</sup>. Например, теоретическая модель научного мышления может включать комбинацию двух умений: формулировать гипотезу и выбирать метод ее проверки.

Рис. 1. Моделирование в психометрике



*Примечание:* Блоки и стрелки — компоненты моделирования, их последовательность и взаимовлияние; штриховой линией помечены гипотетические компоненты. Только затененные компоненты вовлечены в эмпирическую проверку и потому дают легитимные выводы. Выводы из психометрической модели, сделанные в отношении теоретической модели и латентного конструкта, нелегитимны (пунктирные стрелки).

Третий шаг — это проверка теоретической модели. Для этой проверки используется психометрическая модель. Если она является формализованным вариантом психологической теории конструкта, будут сформированы соответствующие статистические ожидания от распределения данных по тесту, например о размерности (факторной структуре) теста и соответствии опре-

<sup>4</sup> Вообще говоря, теории относительно психологических феноменов могут быть вполне сформированными, например теория Ж. Пиаже о развитии мышления ребенка, которая разрабатывалась много лет. Такие теории уже не полагаются на интуиции, а имеют в своей основе систематически собранный экспериментальный или другой эмпирический материал. Однако такие теории нечасто используются для представления «латентных конструктов». Причины этого, а также различия между экспериментальными методами психологии и психометрическим моделированием не будут обсуждаться в этой работе, хотя они чрезвычайно важны.

деленных заданий «своей» размерности. Так, данные по тесту научного мышления в соответствии с вышеуказанной теорией должны образовывать два фактора: включающие задания на построение гипотез и на знание различных методов исследования.

Технически проверка теоретической модели конструкта реализуется через психометрическое описание полученных данных — через проверку размерности, распределения заданий по размерностям и проч. Причем важно, чтобы формально-статистическое описание соответствовало полученным данным как можно более точно. Наилучшая модель затем сравнивается с теоретически ожидаемой.

При выборе наилучшей модели исследователь опирается на результаты анализа согласия модели с полученными данными. Оценка согласия, как правило, включает оценку размерности, согласия данных с моделью на уровне отдельных заданий, согласия модели с данными с принятием решения о модели, лучше всего подходящей данным, а также локальной независимости заданий, предсказаний модели в отношении получаемых данных по тесту и др. [Hambleton, Swaminathan, 2013; Yen, Fitzpatrick, 2006].

Остановимся на философии выбора модели чуть подробнее, потому что это единственный способ принять, отклонить или скорректировать психометрическую модель — а следовательно, и модель теоретическую. В целом выбор модели может быть описан как итеративный процесс, опции которого располагаются в континууме: с одной стороны, модель определяет, какие должны быть данные, чтобы соответствовать модели, так что несогласующиеся данные следует отбросить [Rasch, 1960], а с другой стороны, собранные по тесту данные определяют то, какая модель должна быть найдена, чтобы наилучшим образом согласовываться с ними [Hambleton, Swaminathan, 2013; Yen, Fitzpatrick, 2006]. В 70–80-х годах прошлого века, когда IRT-модели постепенно приобретали популярность, между сторонниками разных подходов к выбору модели шли интенсивные дебаты [Divgi, 1986]. В настоящее время решение о выборе модели может основываться на типе полученных данных (ранги, классификация или их комбинация), целях моделирования (использовать результаты для оценки респондентов, отразить в модели композитность конструкта и проч.), выдвинутых моделью допущений, соображений удобства статистической обработки данных и интерпретации [Luo, 2021; Robitzsch, 2022].

Когда решение принято, исследователь может сделать два класса выводов (рис. 1). Первый касается шкалирования, начисления баллов за тест, выделения в тесте подшкал, надежности инструмента, целесообразности подсчета общего балла по тесту и проч. [Reise, 2012]. В отличие от этих, сугубо практи-

ческих выводов, выводы второго класса имеют непосредственное отношение к латентному конструкту и соответствующей теоретической модели: согласованная с данными психометрическая модель прямо накладывается на структуру латентного конструкта. Иллюстрацией таких выводов служат формулировки, используемые в статьях и их заголовках, например:

«Обзор психометрических подходов к раскрытию структуры психиатрических конструктов» [Borsboom et al., 2016]; «Цель исследования заключалась в анализе психометрических свойств испанской версии шкалы тревожности в связи с коронавирусом (CAS) с использованием <...> IRT и конфирматорного факторного анализа <...> Модели <...> показали, что CAS более информативен при высоком уровне тревожности по COVID-19. CAS обладает достаточными психометрическими свойствами для использования в качестве краткой меры тревожности по COVID-19» [Caycho-Rodríguez et al., 2022]; «Из трех протестированных нами моделей неприемлемой была признана та, которая представляла реляционное мышление как одномерный конструкт <...> Оставшиеся две модели представляли реляционное мышление как многомерное <...> Что в конечном итоге отличало эти две модели, так это включение фактора более высокого порядка сверх факторов, представляющих четыре проявления реляционного мышления» [Alexander et al., 2016].

Однако для такого рода утверждений о конструкте требуется, чтобы, во-первых, тест и данные по тесту были связаны с исследуемым конструктом, и только с ним, и, во-вторых, чтобы была известна функция, связывающая тест и конструкт. Но выполняется ли это условие?

### **Задания теста и конструкт**

Задания теста разрабатываются с таким расчетом, чтобы они вызвали поведение, подпадающее под описание латентного конструкта и/или его компонентов [Messick, 1994; Kane, 2016]. Например, если латентный конструкт — это способность к научному мышлению, а теоретическая модель научного мышления включает два его аспекта, а именно умение формулировать гипотезы и знание методов их проверки, то задания теста будут требовать, соответственно, формулировки гипотезы и указания методов их проверки.

Удостовериться в том, что задания вызывают нужное поведение, можно, проведя предварительные «качественные» исследования. Так, можно убедиться, что задание на формулировку гипотезы провоцирует формулировку гипотезы, а не вспоминание научных фактов, например. Кроме того, в «каче-

ственном» исследовании проверяется, понятны ли вопросы, удобно ли расположен стимульный материал, достаточно ли выделенного на решение времени и проч.<sup>5</sup> [Mislevy, Steinberg, Almond, 2002; Messick, 1994].

Однако из полученных в предварительном исследовании свидетельств того, что задания теста вызывают нужное поведение, не следует, что это поведение детерминировано латентным конструктом. И тем более эти свидетельства не проясняют функцию, связывающую особенности того или иного поведения (например, скорость или правильность решения задачи) с уровнем выраженности латентного конструкта. Причин тому две. Во-первых, различия между людьми в выполнении тех или иных действий могут объясняться не искомым специфическим конструктом, а, например, предыдущим опытом, специальными знаниями, особенностями нервной системы, а также взаимодействием этих факторов [Costantini et al., 2015]. Во-вторых, типичный дизайн исследования, направленного на сбор свидетельств валидности теста, — корреляционный. Иными словами, никакая критериальная валидизация и полученные в результате статистики в принципе не могут гарантировать тесту статус измеряющего инструмента или хотя бы инструмента, чувствительного к искомому конструкту [Borsboom, Mellenbergh, van Heerden, 2004; Trendler, 2013; 2022; Uher, 2021].

Задача представить баллы по тесту как математическую функцию от изучаемого психологического конструкта ( $y = f(x)$ ) принципиально нерешаема ни процедурно (например, экспертизой), ни технологически (например, *evidence-centered design*). Для ее решения должно быть выполнено невыполнимое требование: необходимо иметь одновременные значения для обоих членов уравнения, тогда как мы можем наблюдать значения только для одного из них — это ответ на задание теста. Эта ситуация с тестами — современная версия психофизической проблемы, известной в истории психологии с тех времен, когда психофизики пытались установить соответствие между феноменами «телесными» и феноменами «душевными» [Johnson, 1945]. Все попытки решить эту проблему приводили и приводят только к «предположительным выводам», которые нельзя проверить эмпирически.

Таким образом, каким бы способом ни был разработан тест, он лишь гипотетически соотносится с искомым латентным конструктом. Поэтому тест не может быть использован в моделировании как источник данных о конструкте.

<sup>5</sup> Разработка теста — гораздо более сложный процесс, чем описано, здесь представлены только принципиальные шаги и направления проверки.

**Легитимность выводов психометрического моделирования**

Повторимся: достигнутое согласие модели и данных позволяет делать два класса выводов — прикладные и теоретические. Прикладные выводы касаются структуры данных (например, вывод о количестве факторов) и оптимальных вариантов скоринга. Прикладные выводы легитимны, так как именно структура распределения данных и оценивалась. Теоретические выводы касаются латентного конструкта: его психологической структуры и его оценки. Эти выводы нелегитимны (см. рис. 1, пунктирные стрелки).

Моделирование конструкта в психометрике не следует логике репрезентации конструкта, так как структура последнего неизвестна и поэтому репрезентировать его невозможно. Моделирование конструкта не следует и логике обнаружения объекта, так как для плохо изученного конструкта (феномена) не описан класс объектов, признаки которых нужно обнаружить, не создан протокол обнаружения, не сформированы правила вывода и принятия решения об обнаружении. В отличие от моделей в естественных науках, в психометрике правила связывания входящих переменных с «латентными конструктами» подбираются в процессе анализа и независимо от функционирования изучаемой эмпирической системы, поэтому модель не обнаруживает «латентные переменные», но конструирует их способом, наиболее оптимальным для предписанной модели и входящих переменных [Maraun, Halpin, 2008]. Кроме того, не вводятся ограничения адекватности модели: модель предположительно может адекватно описывать «структуру конструкта» для всех людей, независимо от возраста, рода деятельности, социального и культурного контекста, здоровья и т.д. Так что, какой бы модели в конечном итоге ни соответствовали данные тестирования и как бы модель ни отклонялась от данных, мы не получим информации о конструкте просто потому, что неизвестно, с чем связано отклонение модели или согласие данных с моделью — с данными или с моделью. Любые результаты моделирования можно объяснить и неверным представлением о конструкте (например, двухкомпонентная теория неверна), и неверными данными (например, погрешностями теста или выборки).

**Моделирование и ad hoc теории в психологии**

Прямым следствием такого положения дел является генерация *ad hoc*<sup>6</sup> теорий относительно психологического конструкта (феномена). Их особенность состоит в том, что всякий раз, когда удастся привести к согласию какую-то модель с каким-то набором данных, эта модель принимается как теория, объясняющая

<sup>6</sup> *ad hoc* — теории по случаю, пригодные для отдельного случая.

этот набор данных. При этом, если окажется, что эта модель не воспроизводится на новых данных, она модифицируется с таким расчетом, чтобы лучше подходить для новых данных, и превращается в новую «теорию». При этом старая модель продолжает существовать как объяснение, пригодное для предыдущего случая. То есть если один набор данных согласуется с моделью  $X$ , а другой — с моделью  $Y$ , то обе модели и оба теоретических объяснения валидны. В итоге возникает ситуация, когда в отношении одного и того же конструкта может существовать неограниченное число моделей, приведенных в согласие с данными тестов, так что конструкт может быть описан разными структурами, равно пригодными и равно не опровергнутыми [Vessonen, 2021].

*Ad hoc* теории можно легко обнаружить в рассуждениях о чертах или способностях, когда эти рассуждения опираются на данные моделирования. Например, по мысли авторов идеи конструкта «настойчивость», последняя состоит из «стабильности интересов» и «настойчивости усилий» [Duckworth, Quinn, 2012; Duckworth et al., 2007]. Конструкт исследуется в основном с помощью разных версий шкалы «Настойчивость». Результаты статистического анализа распределения ответов (структуры шкалы) существенно разнятся в разных исследованиях: одномерная, бифакторная или иерархическая структура, зависимы или независимы факторы друг от друга, является ли структура инвариантной и проч. Вследствие прямого накладывания статистической структуры данных теста на латентный конструкт настойчивость описывается соответственно как имеющая бифакторную, иерархическую структуру и т.п. [Tumeneva, Karpapova, Kuzmina, 2019; Credé, 2018; Tynan, 2021]. Однако, поскольку в психометрических исследованиях черта не исследуется, а конструируется в процессе моделирования на основе тестовых данных, нет никаких перспектив опровергнуть или подтвердить эти теории. Теория, сконструированная на базе психометрической модели, всегда будет соответствовать «своим» данным, а новые данные будут находить «свою» модель.

### **Заключительные замечания**

Наш анализ показал, что то, что психометрики имеют в виду под моделированием, не является моделированием в общенаучном понимании этого слова: ни в смысле репрезентации, ни в смысле обнаружения ненаблюдаемого объекта. Поэтому из двух классов выводов, обычно следующих из результатов психометрического моделирования — о структуре тестовых данных и о структуре латентного конструкта, — последние нельзя считать легитимными. Невозможно построить модель ненаблюдаемого объекта (конструкта), природа и закономер-

ности функционирования которого не установлены, полагаясь на данные, связь которых с конструктом также не установлена.

Вместо того чтобы использовать результаты психометрического анализа данных теста строго по назначению — для шкалирования, скоринга, выявления неинформативных заданий и проч., исследователи делают из них содержательные выводы о психологических характеристиках. Такое использование психометрики не только тормозит развитие психологического знания в целом, но и уводит его в сторону безосновательного и непродуктивного «теоретизирования».

Причины распространения практики прямого переноса выводов психометрического моделирования на психологические феномены заслуживают особого внимания, но не будут здесь обсуждаться. Эта работа ставила целью лишь показать, что именно не дает психометрическому моделированию занять нишу полноценного исследовательского инструмента в психологии, каким бы удобным и привычным ни представлялся этот путь.

#### **Благодарности**

Исследование реализовано при поддержке факультета социальных наук, Национальный исследовательский университет «Высшая школа экономики».

Автор благодарит обоих рецензентов за содержательную критику и мотивацию к дискуссии.

#### **Литература**

1. Айзель Г.В., Гусев Е.М., Насонова О.Н. (2017) Расчеты речного стока на основе модели SWAP для водосборов с недостаточным информационным обеспечением. 2. Использование методов физико-географического подобию и пространственной геостатистики. *Водные ресурсы*, т. 44, № 4, сс. 419–431. <https://doi.org/10.7868/S0321059617040022>
2. Угланова И.Л., Брун И.В., Васин Г.М. (2018) Методология *Evidence-Centered Design* для измерения комплексных психологических конструктов. *Современная зарубежная психология*, т. 7, № 3, сс. 18–27. <https://doi.org/10.17759/jmfp.2018070302>
3. Ackerman T.A., Gierl M.J., Walker C.M. (2003) Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, vol. 22, no 3, pp. 37–51. <http://dx.doi.org/10.1111/j.1745-3992.2003.tb00136.x>
4. Alexander P.A., Dumas D., Grossnickle E.M., List A., Firetto C.M. (2016) Measuring Relational Reasoning. *The Journal of Experimental Education*, vol. 84, no 1, pp. 119–151. <http://dx.doi.org/10.1080/00220973.2014.963216>
5. Araujo A.L.S.O., Andrade W.L., Guerrero D.D.S., Melo M.R.A. (2019) How Many Abilities Can We Measure in Computational Thinking? A Study on Bebras Challenge. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, 2019, 27 February), New York, NY: Machinery, pp. 545–551.

6. Arhonditsis G.B., Stow C.A., Steinberg L.J., Kenney M.A., Lathrop R.C., McBride S.J., Reckhow K.H. (2006) Exploring Ecological Patterns with Structural Equation Modeling and Bayesian Analysis. *Ecological Modelling*, vol. 192, no 3–4, pp. 385–409. <https://doi.org/10.1016/j.ecolmodel.2005.07.028>
7. Birenbaum M., DeLuca C., Earl L., Heritage M., Klenowski V., Looney A. et al. (2015) International Trends in the Implementation of Assessment for Learning: Implications for Policy and Practice. *Policy Futures in Education*, vol. 13, no 1, pp. 117–140. <http://dx.doi.org/10.1177/1478210314566733>
8. Birnbaum M.H. (2008) New Paradoxes of Risky Decision Making. *Psychological Review*, vol. 115, no 2, pp. 463–501. <https://doi.org/10.1037/0033-295X.115.2.463>
9. Borsboom D., Mellenbergh G.J., van Heerden J. (2004) The Concept of Validity. *Psychological Review*, vol. 111, no 4, pp. 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
10. Borsboom D., Molenaar D. (2015) Psychometrics. *International Encyclopedia of the Social & Behavioral Sciences* (ed. J.D. Wright), Oxford: Elsevier, pp. 418–422. <https://doi.org/10.1016/B978-0-08-097086-8.43079-5>
11. Borsboom D., Rhemtulla M., Cramer A.O., van der Maas H.L., Scheffer M., Dolan C.V. (2016) Kinds Versus Continua: A Review of Psychometric Approaches to Uncover the Structure of Psychiatric Constructs. *Psychological Medicine*, vol. 46, no 8, pp. 1567–1579. <http://dx.doi.org/10.1017/S0033291715001944>
12. Buchholz J., Hartig J. (2020) Measurement Invariance Testing in Questionnaires: A Comparison of Three Multigroup-CFA and IRT-Based Approaches. *Psychological Test and Assessment Modeling*, vol. 62, no 1, pp. 29–53.
13. Caycho-Rodríguez T., Vilca L.W., Carbajal-León C., White M., Vivanco-Vidal A., Saroli-Araníbar D. et al. (2022) Coronavirus Anxiety Scale: New Psychometric Evidence for the Spanish Version Based on CFA and IRT Models in a Peruvian Sample. *Death Studies*, vol. 46, no 5, pp. 1090–1099. <http://dx.doi.org/10.1080/07481187.2020.1865480>
14. Costantini G., Epskamp S., Borsboom D., Perugini M., Möttus R., Waldorp L.J., Cramer A.O. (2015) State of the aRt Personality Research: A Tutorial on Network Analysis of Personality Data in R. *Journal of Research in Personality*, vol. 54, July, pp. 13–29. <https://doi.org/10.1016/j.jrp.2014.07.003>
15. Credé M. (2018) What Shall We Do about Grit? A Critical Review of What We Know and What We Don't Know. *Educational Researcher*, vol. 47, no 9, pp. 606–611. <http://dx.doi.org/10.3102/0013189X18801322>
16. Cronbach L.J., Meehl P.E. (1955) Construct Validity in Psychological Tests. *Psychological Bulletin*, vol. 52, no 4, pp. 281–302. <https://doi.org/10.1037/h0040957>
17. Dam van J.C., Groenendijk P., Hendriks R.F., Kroes J.G. (2008) Advances of Modeling Water Flow in Variably Saturated Soils with SWAP. *Vadose Zone Journal*, vol. 7, no 2, pp. 640–653. <http://dx.doi.org/10.2136/vzj2007.0060>
18. Divgi D.R. (1986) Does the Rasch Model Really Work for Multiple Choice Items? Not If You Look Closely. *Journal of Educational Measurement*, vol. 23, no 4, pp. 283–298.
19. Duckworth A.L., Quinn P.D. (2012) Short Grit Scale. *Journal of Personality Assessment*, vol. 91, no 2, pp. 166–174. <https://psycnet.apa.org/doi/10.1037/t01598-000>
20. Duckworth A.L., Peterson C., Matthews M.D., Kelly D.R. (2007) Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*, vol. 92, no 6, 1087–1101. <http://dx.doi.org/10.1037/0022-3514.92.6.1087>
21. Dumas D., Dong Y. (2022) Relational Reasoning and Thinking: Theory, Measurement, and Empirical Findings. *International Encyclopedia of Education* (eds R. Tierney, F. Rizvi, K. Ercican), New York, NY: Taylor & Francis. <https://doi.org/10.4324/9781138609877-REE179-1>

22. Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. [http://dx.doi.org/10.1016/0001-6918\(73\)90003-6](http://dx.doi.org/10.1016/0001-6918(73)90003-6)
23. Fisher Jr. W.P., Stenner A.J. (2022) Metrology for the Social, Behavioral, and Economic Sciences. *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (eds W.P. Fisher, P.J. Massengill), Singapore: Springer Nature Singapore, pp. 217–222.
24. Fox J.P. (2005) Multilevel IRT Using Dichotomous and Polytomous Response Data. *British Journal of Mathematical and Statistical Psychology*, vol. 58, no 1, pp. 145–172. <http://dx.doi.org/10.1348/000711005X38951>
25. Franić S., Borsboom D., Dolan C.V., Boomsma D.I. (2014) The Big Five Personality Traits: Psychological Entities or Statistical Constructs? *Behavior Genetics*, vol. 44, no 6, pp. 591–604. <http://dx.doi.org/10.1007/s10519-013-9625-7>
26. Franic S., Dolan C.V., Borsboom D., Boomsma D.I. (2012) Structural Equation Modeling in Genetics. *Handbook of Structural Equation Modeling* (ed. R.H. Hoyle), New York, NY: The Guilford, pp. 617–635.
27. Freund P.A., Lohbeck A. (2021) Modeling Self-Determination Theory Motivation Data by Using Unfolding IRT. *European Journal of Psychological Assessment*, vol. 37, no 5, pp. 388–396. <http://dx.doi.org/10.1027/1015-5759/a000629>
28. Hambleton R.K., Swaminathan H. (2013) *Item Response Theory: Principles and Applications*. Springer Science & Business Media.
29. Hartig J., Höhler J. (2009) Multidimensional IRT Models for the Assessment of Competencies. *Studies in Educational Evaluation*, vol. 35, no 2–3, pp. 57–63. <http://dx.doi.org/10.1016/j.stueduc.2009.10.002>
30. Hauwaert van S.M., Schimpf C.H., Azevedo F. (2020) The Measurement of Populist Attitudes: Testing Cross-National Scales Using Item Response Theory. *Politics*, vol. 40, no 1, Article no 026339571985930. <http://dx.doi.org/10.1177/0263395719859306>
31. Johnson H.M. (1945) Are Psychophysical Problems Genuine or Spurious? *The American Journal of Psychology*, vol. 58, no 2, pp. 189–211. <https://doi.org/10.2307/1417845>
32. Kane M.T. (2016) Explicating Validity. *Assessment in Education: Principles, Policy & Practice*, vol. 23, no 2, pp. 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
33. Kunina-Habenicht O., Goldhammer F. (2020) ICT Engagement: A New Construct and Its Assessment in PISA 2015. *Large-Scale Assessments in Education*, vol. 8, no 1, pp. 1–21. <http://dx.doi.org/10.1186/s40536-020-00084-z>
34. Lange J., Dalege J., Borsboom D., van Kleef G.A., Fischer A.H. (2020) Toward an Integrative Psychometric Model of Emotions. *Perspectives on Psychological Science*, vol. 15, no 2, pp. 444–468. <http://dx.doi.org/10.1177/1745691619895057>
35. Linden van der W.J., Hambleton R.K. (eds) (2013) *Handbook of Modern Item Response Theory*. Springer Science & Business Media.
36. Luo Y. (2021) A Comparison of Common IRT Model-selection Methods with Mixed-Format Tests. *Measurement: Interdisciplinary Research and Perspectives*, vol. 19, no 4, pp. 199–212. <http://dx.doi.org/10.1080/15366367.2021.1878779>
37. MacCorquodale K., Meehl P.E. (1948) On a Distinction between Hypothetical Constructs and Intervening Variables. *Psychological Review*, vol. 55, no 2, pp. 95–107. <https://doi.org/10.1037/h0056029>
38. Maraun M. (2017) The Object Detection Logic of Latent Variable Technologies. *Quality and Quantity*, vol. 51, no 1, pp. 239–259. <https://doi.org/10.1007/s11135-015-0303-0>
39. Maraun M.D., Gabriel S.M. (2013) Illegitimate Concept Equating in the Partial Fusion of Construct Validation Theory and Latent Variable Modeling. *New Ideas in Psychology*, vol. 31, no 1, pp. 32–42. <https://doi.org/10.1016/j.newideapsych.2011.02.006>

40. Maraun M.D., Halpin P.F. (2008) Manifest and Latent Variates. *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no 1–2, pp. 113–117. <https://doi.org/doi:10.1080/15366360802035596>
41. Markus K.A., Borsboom D. (2013) *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York, NY: Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9780203501207>
42. Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. <https://doi.org/10.3102/0013189x023002013>
43. Michell J. (2013) Constructs, Inferences, and Mental Measurement. *New Ideas in Psychology*, vol. 31, no 1, pp. 13–21. <https://doi.org/10.1016/j.newideapsych.2011.02.004>
44. Mislevy R.J., Steinberg L.S., Almond R.G. (2002) On the Roles of Task Model Variables in Assessment Design. *Generating Items for Cognitive Tests: Theory and Practice* (eds S. Irvine, P. Kyllonen), Hillsdale, NY: Erlbaum, pp. 97–128.
45. Nering M.L., Ostini R. (eds) (2010) *Handbook of Polytomous Item Response Theory Models*. New York, NY: Routledge. <https://doi.org/10.4324/9780203861264>
46. Nima A.A., Cloninger K.M., Persson B.N., Sikström S., Garcia D. (2020) Validation of Subjective Well-Being Measures Using Item Response Theory. *Frontiers in Psychology*, vol. 10, January, Article no 3036. <http://dx.doi.org/10.3389/fpsyg.2019.03036>
47. Oberkampf W.L., DeLand S.M., Rutherford B.M., Diegert K.V., Alvin K.F. (2002) Error and Uncertainty in Modeling and Simulation. *Reliability Engineering & System Safety*, vol. 75, no (3), pp. 333–357. [http://dx.doi.org/10.1016/S0951-8320\(01\)00120-X](http://dx.doi.org/10.1016/S0951-8320(01)00120-X)
48. Ottensen J. (2000) *Mathematical Modelling in Medicine*. Amsterdam: IOS Press.
49. Podolsky A., Kaufman K.R., Cahalan T.D., Aleshinsky S.Y., Chao E.Y. (1990) The Relationship of Strength and Jump Height in Figure Skaters. *The American Journal of Sports Medicine*, vol. 18, no 4, pp. 400–405. <https://doi.org/10.1177/036354659001800412>
50. Power M.J. (2006) The Structure of Emotion: An Empirical Comparison of Six Models. *Cognition and Emotion*, vol. 20, no 5, pp. 694–713. <https://doi.org/10.1080/02699930500367925>
51. Pugesek B.H., Tomer A., von Eye A. (2003) *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*. Cambridge, UK: Cambridge University. <https://doi.org/10.1017/CBO9780511542138>
52. Qian M., Plucker J.A., Yang X. (2019) Is Creativity Domain Specific or Domain General? Evidence from Multilevel Explanatory Item Response Theory Models. *Thinking Skills and Creativity*, vol. 33, May, Article no 100571. <http://dx.doi.org/10.1016/j.tsc.2019.100571>
53. Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
54. Ravand H., Robitzsch A. (2015) Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research, and Evaluation*, vol. 20, no 11. Available at: <http://pareonline.net/getvn.asp?v=20&n=11> (accessed 20 August 2023).
55. Reise S.P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, vol. 47, no 5, pp. 667–696. <https://doi.org/10.1080/00273171.2012.715555>
56. Rhodes M., Putkaradze V. (2022) Trajectory Tracing in Figure Skating. *Nonlinear Dynamics*, vol. 110, no 4, pp. 3031–3044. <https://doi.org/10.1007/s11071-022-07806-8>
57. Riconscente M.M., Mislevy R.J., Corrigan S. (2015) Evidence-Centered Design. *Handbook of Test Development* (eds S. Lane, M.R. Raymond, T.M. Haladyna), New York, NY: Routledge, pp. 40–63. <http://dx.doi.org/10.4324/9780203102961.ch3>

58. Robitzsch A. (2022) On the Choice of the Item Response Model for Scaling PISA Data: Model Selection Based on Information Criteria and Quantifying Model Uncertainty. *Entropy*, vol. 24, no 6, Article no 760. <http://dx.doi.org/10.3390/e24060760>
59. Schmittmann V.D., Cramer A.O.J., Waldorp L.J., Epskamp S., Kievit R.A., Borsboom D. (2013) Deconstructing the Construct: A Network Perspective on Psychological Phenomena. *New Ideas in Psychology*, vol. 31, no 1, pp. 43–53. <https://doi.org/10.1016/j.newideapsych.2011.02.007>
60. Sen S., Cohen A.S. (2019) Applications of Mixture IRT Models: A Literature Review. *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no 4, pp. 177–191. <http://dx.doi.org/10.1080/15366367.2019.1583506>
61. Shaw A., Kapnek M., Morelli N.A. (2021) Measuring Creative Self-Efficacy: An Item Response Theory Analysis of the Creative Self-Efficacy Scale. *Frontiers in Psychology*, vol. 12, July, Article no 678033. <http://dx.doi.org/10.3389/fpsyg.2021.678033>
62. Sijtsma K., Ark van der A. (2020) *Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models*. Boca Raton, FL: CRC. <https://doi.org/10.1201/9780429112447>
63. Streckert N., Kurtz L., Kajonius P.J. (2023) Can Your Darkness Be Measured? Analyzing the Full and Brief Version of the Dark Factor of Personality in Swedish. *International Journal of Testing*, vol. 23, no 2, pp. 1–45. <http://dx.doi.org/10.1080/15305058.2023.2195659>
64. Templin J.L., Henson R.A. (2006) Measurement of Psychological Disorders Using Cognitive Diagnosis Models. *Psychological Methods*, vol. 11, no 3, pp. 287–305. <http://dx.doi.org/10.1037/1082-989X.11.3.287>
65. Trendler G. (2022) Is Measurement in Psychology an Empirical or a Conceptual Issue? A Comment on David Franz. *Theory & Psychology*, vol. 32, no 1, pp. 164–170. <https://doi.org/10.1177/09593543211050025>
66. Trendler G. (2013) Measurement in Psychology: A Case of Ignoramus et Ignorabimus? A Rejoinder. *Theory & Psychology*, vol. 23, no 5, pp. 591–615. <https://doi.org/10.1177/0959354313490451>
67. Tynan M.C. (2021) Deconstructing Grit's Validity: The Case for Revising Grit Measures and Theory. *Multidisciplinary Perspectives on Grit: Contemporary Theories, Assessments, Applications and Critiques* (eds L.E. van Zyl, C. Olckers, L. van der Vaart), Cham: Springer Nature Switzerland, pp. 137–155. [http://dx.doi.org/10.1007/978-3-030-57389-8\\_8](http://dx.doi.org/10.1007/978-3-030-57389-8_8)
68. Tyumeneva Y., Kardanova E., Kuzmina J. (2019) Grit: Two Related but Independent Constructs Instead of One. Evidence from Item Response Theory. *European Journal of Psychological Assessment*, vol. 35, no 4, pp. 469–478. <http://dx.doi.org/10.1027/1015-5759/a000424>
69. Uher J. (2021) Quantitative Psychology under Scrutiny: Measurement Requires Not Result-Dependent But Traceable Data Generation. *Personality and Individual Differences*, vol. 170, no 5, Article no 110205. <https://doi.org/10.1016/j.paid.2020.110205>
70. Vessonen E. (2021) Conceptual Engineering and Operationalism in Psychology. *Synthese*, vol. 199, no 3–4, pp. 10615–10637. <https://doi.org/10.1007/s11229-021-03261-x>
71. Wagner T.A., Harvey R.J. (2006) Development of a New Critical Thinking Test Using Item Response Theory. *Psychological Assessment*, vol. 18, no 1, pp. 100–105. <https://doi.org/10.1037/1040-3590.18.1.100>
72. Walton K.E., Roberts B.W., Krueger R.F., Blonigen D.M., Hicks B.M. (2008) Capturing Abnormal Personality with Normal Personality Inventories: An Item Response Theory Approach. *Journal of Personality*, vol. 76, no 6, pp. 1623–1648. <http://dx.doi.org/10.1111/j.1467-6494.2008.00533.x>

73. Wiggins B.J., Christopherson C.D. (2019) The Replication Crisis in Psychology: An Overview for Theoretical and Philosophical Psychology. *Journal of Theoretical and Philosophical Psychology*, vol. 39, no 4, pp. 202–217. <http://dx.doi.org/10.1037/teo0000137>
74. Will C.M. (2000) *Einstein's Relativity and Everyday Life*. Available at: <http://www.physicscentral.com/writers/writers-00-2.html> (accessed 20 August 2023).
75. Wilson M. (2004) *Constructing Measures. An Item Response Modeling Approach*. New York, NY: Routledge.
76. Yen W.M., Fizepatrick A.R. (2006) Item Response Theory. *Educational Measurement* (ed. R.L. Brennan), Westport, CT: American Council on Education and Praeger, pp. 17–64.
77. Zhao H., Alexander P.A., Sun Y. (2021) Relational Reasoning's Contributions to Mathematical Thinking and Performance in Chinese Elementary and Middle-School Students. *Journal of Educational Psychology*, vol. 113, no 2, pp. 279–303. <http://dx.doi.org/10.1037/edu0000595>

## References

- Ackerman T.A., Gierl M.J., Walker C.M. (2003) Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, vol. 22, no 3, pp. 37–51. <http://dx.doi.org/10.1111/j.1745-3992.2003.tb00136.x>
- Alexander P.A., Dumas D., Grossnickle E.M., List A., Firetto C.M. (2016) Measuring Relational Reasoning. *The Journal of Experimental Education*, vol. 84, no 1, pp. 119–151. <http://dx.doi.org/10.1080/00220973.2014.963216>
- Araujo A.L.S.O., Andrade W.L., Guerrero D.D.S., Melo M.R.A. (2019) How Many Abilities Can We Measure in Computational Thinking? A Study on Bebras Challenge. Proceedings of the 50th ACM Technical Symposium on Computer Science Education (Minneapolis, MN, 2019, 27 February), New York, NY: Machinery, pp. 545–551.
- Arhonditsis G.B., Stow C.A., Steinberg L.J., Kenney M.A., Lathrop R.C., McBride S.J., Reckhow K.H. (2006) Exploring Ecological Patterns with Structural Equation Modeling and Bayesian Analysis. *Ecological Modelling*, vol. 192, no 3–4, pp. 385–409. <https://doi.org/10.1016/j.ecolmodel.2005.07.028>
- Ayzel G.V., Gusev E.M., Nasonova O.N. (2017) Raschetnyy rezhim stoka na osnove modeli SWAP dlya vodosborov s nedostatochnym informatsionnym obespecheniem. 2. Ispol'zovanie metodov fiziko-geograficheskogo podpbiya i prostanstvennoy geostatistiki [Runoff Evaluation for Ungauged Watersheds by SWAP Model. 2. Using Methods of Physical and Geographical Similarity and Spatial Geostatistics]. *Water Resources*, vol. 44, no 4, pp. 419–431. <https://doi.org/10.7868/S0321059617040022>
- Birenbaum M., DeLuca C., Earl L., Heritage M., Klenowski V., Looney A. et al. (2015) International Trends in the Implementation of Assessment for Learning: Implications for Policy and Practice. *Policy Futures in Education*, vol. 13, no 1, pp. 117–140. <http://dx.doi.org/10.1177/1478210314566733>
- Birnbaum M.H. (2008) New Paradoxes of Risky Decision Making. *Psychological Review*, vol. 115, no 2, pp. 463–501. <https://doi.org/10.1037/0033-295X.115.2.463>
- Borsboom D., Mellenbergh G.J., van Heerden J. (2004) The Concept of Validity. *Psychological Review*, vol. 111, no 4, pp. 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Borsboom D., Molenaar D. (2015) Psychometrics. *International Encyclopedia of the Social & Behavioral Sciences* (ed. J.D. Wright), Oxford: Elsevier, pp. 418–422. <https://doi.org/10.1016/B978-0-08-097086-8.43079-5>
- Borsboom D., Rhemtulla M., Cramer A.O., van der Maas H.L., Scheffer M., Dolan C.V. (2016) Kinds Versus Continua: A Review of Psychometric Approaches to Un-

- cover the Structure of Psychiatric Constructs. *Psychological Medicine*, vol. 46, no 8, pp. 1567–1579. <http://dx.doi.org/10.1017/S0033291715001944>
- Buchholz J., Hartig J. (2020) Measurement Invariance Testing in Questionnaires: A Comparison of Three Multigroup-CFA and IRT-Based Approaches. *Psychological Test and Assessment Modeling*, vol. 62, no 1, pp. 29–53.
- Caycho-Rodríguez T., Vilca L.W., Carbajal-León C., White M., Vivanco-Vidal A., Saroli-Aranibar D. et al. (2022) Coronavirus Anxiety Scale: New Psychometric Evidence for the Spanish Version Based on CFA and IRT Models in a Peruvian Sample. *Death Studies*, vol. 46, no 5, pp. 1090–1099. <http://dx.doi.org/10.1080/07481187.2020.1865480>
- Costantini G., Epskamp S., Borsboom D., Perugini M., Möttus R., Waldorp L.J., Cramer A.O. (2015) State of the aRt Personality Research: A Tutorial on Network Analysis of Personality Data in R. *Journal of Research in Personality*, vol. 54, July, pp. 13–29. <https://doi.org/10.1016/j.jrp.2014.07.003>
- Credé M. (2018) What Shall We Do about Grit? A Critical Review of What We Know and What We Don't Know. *Educational Researcher*, vol. 47, no 9, pp. 606–611. <http://dx.doi.org/10.3102/0013189X18801322>
- Cronbach L.J., Meehl P.E. (1955) Construct Validity in Psychological Tests. *Psychological Bulletin*, vol. 52, no 4, pp. 281–302. <https://doi.org/10.1037/h0040957>
- Dam van J.C., Groenendijk P., Hendriks R.F., Kroes J.G. (2008) Advances of Modeling Water Flow in Variably Saturated Soils with SWAP. *Vadose Zone Journal*, vol. 7, no 2, pp. 640–653. <http://dx.doi.org/10.2136/vzj2007.0060>
- Divgi D.R. (1986) Does the Rasch Model Really Work for Multiple Choice Items? Not If You Look Closely. *Journal of Educational Measurement*, vol. 23, no 4, pp. 283–298.
- Duckworth A.L., Quinn P.D. (2012) Short Grit Scale. *Journal of Personality Assessment*, vol. 91, no 2, pp. 166–174. <https://psycnet.apa.org/doi/10.1037/t01598-000>
- Duckworth A.L., Peterson C., Matthews M.D., Kelly D.R. (2007) Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*, vol. 92, no 6, 1087–1101. <http://dx.doi.org/10.1037/0022-3514.92.6.1087>
- Dumas D., Dong Y. (2022) Relational Reasoning and Thinking: Theory, Measurement, and Empirical Findings. *International Encyclopedia of Education* (eds R. Tierney, F. Rizvi, K. Ercican), New York, NY: Taylor & Francis. <https://doi.org/10.4324/9781138609877-REE179-1>
- Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. [http://dx.doi.org/10.1016/0001-6918\(73\)90003-6](http://dx.doi.org/10.1016/0001-6918(73)90003-6)
- Fisher Jr. W.P., Stenner A.J. (2022) Metrology for the Social, Behavioral, and Economic Sciences. *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (eds W.P. Fisher, P.J. Massengill), Singapore: Springer Nature Singapore, pp. 217–222.
- Fox J.P. (2005) Multilevel IRT Using Dichotomous and Polytomous Response Data. *British Journal of Mathematical and Statistical Psychology*, vol. 58, no 1, pp. 145–172. <http://dx.doi.org/10.1348/000711005X38951>
- Franić S., Borsboom D., Dolan C.V., Boomsma D.I. (2014) The Big Five Personality Traits: Psychological Entities or Statistical Constructs? *Behavior Genetics*, vol. 44, no 6, pp. 591–604. <http://dx.doi.org/10.1007/s10519-013-9625-7>
- Francis S., Dolan C.V., Borsboom D., Boomsma D.I. (2012) Structural Equation Modeling in Genetics. *Handbook of Structural Equation Modeling* (ed. R.H. Hoyle), New York, NY: The Guilford, pp. 617–635.
- Freund P.A., Lohbeck A. (2021) Modeling Self-Determination Theory Motivation Data by Using Unfolding IRT. *European Journal of Psychological Assessment*, vol. 37, no 5, pp. 388–396. <http://dx.doi.org/10.1027/1015-5759/a000629>
- Hambleton R.K., Swaminathan H. (2013) *Item Response Theory: Principles and Applications*. Springer Science & Business Media.

- Hartig J., Höhler J. (2009) Multidimensional IRT Models for the Assessment of Competencies. *Studies in Educational Evaluation*, vol. 35, no 2–3, pp. 57–63. <http://dx.doi.org/10.1016/j.stueduc.2009.10.002>
- Hauwaert van S.M., Schimpf C.H., Azevedo F. (2020) The Measurement of Populist Attitudes: Testing Cross-National Scales Using Item Response Theory. *Politics*, vol. 40, no 1, Article no 026339571985930. <http://dx.doi.org/10.1177/0263395719859306>
- Johnson H.M. (1945) Are Psychophysical Problems Genuine or Spurious? *The American Journal of Psychology*, vol. 58, no 2, pp. 189–211. <https://doi.org/10.2307/1417845>
- Kane M.T. (2016) Explicating Validity. *Assessment in Education: Principles, Policy & Practice*, vol. 23, no 2, pp. 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kunina-Habenicht O., Goldhammer F. (2020) ICT Engagement: A New Construct and Its Assessment in PISA 2015. *Large-Scale Assessments in Education*, vol. 8, no 1, pp. 1–21. <http://dx.doi.org/10.1186/s40536-020-00084-z>
- Lange J., Dalege J., Borsboom D., van Kleef G.A., Fischer A.H. (2020) Toward an Integrative Psychometric Model of Emotions. *Perspectives on Psychological Science*, vol. 15, no 2, pp. 444–468. <http://dx.doi.org/10.1177/1745691619895057>
- Linden van der W.J., Hambleton R.K. (eds) (2013) *Handbook of Modern Item Response Theory*. Springer Science & Business Media.
- Luo Y. (2021) A Comparison of Common IRT Model-selection Methods with Mixed-Format Tests. *Measurement: Interdisciplinary Research and Perspectives*, vol. 19, no 4, pp. 199–212. <http://dx.doi.org/10.1080/15366367.2021.1878779>
- MacCorquodale K., Meehl P.E. (1948) On a Distinction between Hypothetical Constructs and Intervening Variables. *Psychological Review*, vol. 55, no 2, pp. 95–107. <https://doi.org/10.1037/h0056029>
- Maraun M. (2017) The Object Detection Logic of Latent Variable Technologies. *Quality and Quantity*, vol. 51, no 1, pp. 239–259. <https://doi.org/10.1007/s11135-015-0303-0>
- Maraun M.D., Gabriel S.M. (2013) Illegitimate Concept Equating in the Partial Fusion of Construct Validation Theory and Latent Variable Modeling. *New Ideas in Psychology*, vol. 31, no 1, pp. 32–42. <https://doi.org/10.1016/j.newideapsych.2011.02.006>
- Maraun M.D., Halpin P.F. (2008) Manifest and Latent Variates. *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no 1–2, pp. 113–117. <https://doi.org/10.1080/15366360802035596>
- Markus K.A., Borsboom D. (2013) *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York, NY: Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9780203501207>
- Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. <https://doi.org/10.3102/0013189x023002013>
- Michell J. (2013) Constructs, Inferences, and Mental Measurement. *New Ideas in Psychology*, vol. 31, no 1, pp. 13–21. <https://doi.org/10.1016/j.newideapsych.2011.02.004>
- Mislevy R.J., Steinberg L.S., Almond R.G. (2002) On the Roles of Task Model Variables in Assessment Design. Generating Items for Cognitive Tests: Theory and Practice (eds S. Irvine, P. Kyllonen), Hillsdale, NY: Erlbaum, pp. 97–128.
- Nering M.L., Ostini R. (eds) (2010) *Handbook of Polytomous Item Response Theory Models*. New York, NY: Routledge. <https://doi.org/10.4324/9780203861264>
- Nima A.A., Cloninger K.M., Persson B.N., Sikström S., Garcia D. (2020) Validation of Subjective Well-Being Measures Using Item Response Theory. *Frontiers*

- in Psychology*, vol. 10, January, Article no 3036. <http://dx.doi.org/10.3389/fpsyg.2019.03036>
- Ottensen J. (2000) *Mathematical Modelling in Medicine*. Amsterdam: IOS Press.
- Podolsky A., Kaufman K.R., Cahalan T.D., Aleshinsky S.Y., Chao E.Y. (1990) The Relationship of Strength and Jump Height in Figure Skaters. *The American Journal of Sports Medicine*, vol. 18, no 4, pp. 400–405. <https://doi.org/10.1177/036354659001800412>
- Power M.J. (2006) The Structure of Emotion: An Empirical Comparison of Six Models. *Cognition and Emotion*, vol. 20, no 5, pp. 694–713. <https://doi.org/10.1080/02699930500367925>
- Pugesek B.H., Tomer A., von Eye A. (2003) *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*. Cambridge, UK: Cambridge University. <https://doi.org/10.1017/CBO9780511542138>
- Qian M., Plucker J.A., Yang X. (2019) Is Creativity Domain Specific or Domain General? Evidence from Multilevel Explanatory Item Response Theory Models. *Thinking Skills and Creativity*, vol. 33, May, Article no 100571. <http://dx.doi.org/10.1016/j.tsc.2019.100571>
- Oberkampf W.L., DeLand S.M., Rutherford B.M., Diegert K.V., Alvin K.F. (2002) Error and Uncertainty in Modeling and Simulation. *Reliability Engineering & System Safety*, vol. 75, no 3, pp. 333–357. [http://dx.doi.org/10.1016/S0951-8320\(01\)00120-X](http://dx.doi.org/10.1016/S0951-8320(01)00120-X)
- Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
- Ravand H., Robitzsch A. (2015) Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research, and Evaluation*, vol. 20, no 11. Available at: <http://pareonline.net/getvn.asp?v=20&n=11> (accessed 20 August 2023).
- Reise S.P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, vol. 47, no 5, pp. 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Riconscente M.M., Mislevy R.J., Corrigan S. (2015) Evidence-Centered Design. *Handbook of Test Development* (eds S. Lane, M.R. Raymond, T.M. Haladyna), New York, NY: Routledge, pp. 40–63. <http://dx.doi.org/10.4324/9780203102961.ch3>
- Robitzsch A. (2022) On the Choice of the Item Response Model for Scaling PISA Data: Model Selection Based on Information Criteria and Quantifying Model Uncertainty. *Entropy*, vol. 24, no 6, Article no 760. <http://dx.doi.org/10.3390/e24060760>
- Rhodes M., Putkaradze V. (2022) Trajectory Tracing in Figure Skating. *Nonlinear Dynamics*, vol. 110, no 4, pp. 3031–3044. <https://doi.org/10.1007/s11071-022-07806-8>
- Schmittmann V.D., Cramer A.O.J., Waldorp L.J., Epskamp S., Kievit R.A., Borsboom D. (2013) Deconstructing the Construct: A Network Perspective on Psychological Phenomena. *New Ideas in Psychology*, vol. 31, no 1, pp. 43–53. <https://doi.org/10.1016/j.newideapsych.2011.02.007>
- Sen S., Cohen A.S. (2019) Applications of Mixture IRT Models: A Literature Review. *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no 4, pp. 177–191. <http://dx.doi.org/10.1080/15366367.2019.1583506>
- Shaw A., Kapnek M., Morelli N.A. (2021) Measuring Creative Self-Efficacy: An Item Response Theory Analysis of the Creative Self-Efficacy Scale. *Frontiers in Psychology*, vol. 12, July, Article no 678033. <http://dx.doi.org/10.3389/fpsyg.2021.678033>
- Sijtsma K., Ark van der A. (2020) *Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models*. Boca Raton, FL: CRC. <https://doi.org/10.1201/9780429112447>
- Strecker N., Kurtz L., Kajonius P.J. (2023) Can Your Darkness Be Measured? Analyzing the Full and Brief Version of the Dark Factor of Personality in Swed-

- ish. *International Journal of Testing*, vol. 23, no 2, pp. 1–45. <http://dx.doi.org/10.1080/15305058.2023.2195659>
- Templin J.L., Henson R.A. (2006) Measurement of Psychological Disorders Using Cognitive Diagnosis Models. *Psychological Methods*, vol. 11, no 3, pp. 287–305. <http://dx.doi.org/10.1037/1082-989X.11.3.287>
- Trendler G. (2022) Is Measurement in Psychology an Empirical or a Conceptual Issue? A Comment on David Franz. *Theory & Psychology*, vol. 32, no 1, pp. 164–170. <https://doi.org/10.1177/09593543211050025>
- Trendler G. (2013) Measurement in Psychology: A Case of Ignoramus et Ignorabimus? A Rejoinder. *Theory & Psychology*, vol. 23, no 5, pp. 591–615. <https://doi.org/10.1177/0959354313490451>
- Tyan M.C. (2021) Deconstructing Grit's Validity: The Case for Revising Grit Measures and Theory. *Multidisciplinary Perspectives on Grit: Contemporary Theories, Assessments, Applications and Critiques* (eds L.E. van Zyl, C. Olckers, L. van der Vaart), Cham: Springer Nature Switzerland, pp. 137–155. [http://dx.doi.org/10.1007/978-3-030-57389-8\\_8](http://dx.doi.org/10.1007/978-3-030-57389-8_8)
- Tyumeneva Y., Kardanova E., Kuzmina J. (2019) Grit: Two Related but Independent Constructs Instead of One. Evidence from Item Response Theory. *European Journal of Psychological Assessment*, vol. 35, no 4, pp. 469–478. <http://dx.doi.org/10.1027/1015-5759/a000424>
- Uglanova I.L., Brun I.V., Vasin G.M. (2018) Metodologiya Evidence-Centered Design dlya izmereniya kompleksnykh psikhologicheskikh konstruktov [Evidence-Centered Design Method for Measuring Complex Psychological Constructs]. *Journal of Modern Foreign Psychology*, vol. 7, no 3, pp. 18–27. <https://doi.org/10.17759/jmfp.2018070302>
- Uher J. (2021) Quantitative Psychology under Scrutiny: Measurement Requires Not Result-Dependent But Traceable Data Generation. *Personality and Individual Differences*, vol. 170, no 5, Article no 110205. <https://doi.org/10.1016/j.paid.2020.110205>
- Vessonen E. (2021) Conceptual Engineering and Operationalism in Psychology. *Synthese*, vol. 199, no 3–4, pp. 10615–10637. <https://doi.org/10.1007/s11229-021-03261-x>
- Wagner T.A., Harvey R.J. (2006) Development of a New Critical Thinking Test Using Item Response Theory. *Psychological Assessment*, vol. 18, no 1, pp. 100–105. <https://doi.org/10.1037/1040-3590.18.1.100>
- Walton K.E., Roberts B.W., Krueger R.F., Blonigen D.M., Hicks B.M. (2008) Capturing Abnormal Personality with Normal Personality Inventories: An Item Response Theory Approach. *Journal of Personality*, vol. 76, no 6, pp. 1623–1648. <http://dx.doi.org/10.1111/j.1467-6494.2008.00533.x>
- Wiggins B.J., Christopherson C.D. (2019) The Replication Crisis in Psychology: An Overview for Theoretical and Philosophical Psychology. *Journal of Theoretical and Philosophical Psychology*, vol. 39, no 4, pp. 202–217. <http://dx.doi.org/10.1037/teo0000137>
- Will C.M. (2000) *Einstein's Relativity and Everyday Life*. Available at: <http://www.physiccentral.com/writers/writers-00-2.html> (accessed 20 August 2023).
- Wilson M. (2004) *Constructing Measures. An Item Response Modeling Approach*. New York, NY: Routledge.
- Yen W.M., Fitzpatrick A.R. (2006) Item Response Theory. *Educational Measurement* (ed. R.L. Brennan), Westport, CT: American Council on Education and Praeger, pp. 17–64.
- Zhao H., Alexander P.A., Sun Y. (2021) Relational Reasoning's Contributions to Mathematical Thinking and Performance in Chinese Elementary and Middle-School Students. *Journal of Educational Psychology*, vol. 113, no 2, pp. 279–303. <http://dx.doi.org/10.1037/edu0000595>