Психометрика и когнитивные исследования: противоречия и возможности кооперации

Юлия Кузьмина

Статья поступила в редакцию в марте 2023 г.

Кузьмина Юлия Владимировна — кандидат психологических наук, старший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: papushka7@gmail.com. ORCID: https://orcid.org/0000-0002-4243-8313

Аннотация

Когнитивная психология в основном развивалась в рамках экспериментальной парадигмы, в отличие от психометрики, занимающейся оценкой индивидуальных различий и корреляционными исследованиями. С целью обнаружить барьеры, стоящие на пути сотрудничества когнитивной психологии с психометрикой, в статье рассмотрена краткая история взаимоотношений между экспериментальными исследованиями и психометрикой с конца XIX в. до настоящего времени. Обсуждаются основные проблемы, возникающие в когнитивных исследованиях из-за недостаточного использования психометрических моделей и применения устаревших методов анализа результатов тестирования. По итогам предлагается ряд рекомендаций с точки зрения психометрики для повышения точности измерений индивидуальных различий в когнитивных процессах и способностях.

Ключевые слова

психометрика, когнитивная психология, экспериментальная психология, надежность, индивидуальные различия, анализ времени ответа, ингибиторная функция

Для цитирования

Кузьмина Ю.В. (2023) Психометрика и когнитивные исследования: противоречия и возможности кооперации. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 113–144. https://doi.org/10.17323/vo-2023-16875

Psychometrics and Cognitive Research: Contradictions and Possibility for Cooperation

Yulia Kuzmina

Yulia V. Kuzmina — PhD in Psychology, Researcher at the Center for Psychometrics and Measurement in Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: papushka7@gmail.com. ORCID: https://orcid.org/0000-0002-4243-8313

Abstract The article considered several issues of the relationships between cognitive psychology and psychometrics. Cognitive psychology has mainly developed within experimental paradigm in psychology, whereas psychometrics has developed within a different paradigm — assessment of individual differences and correlational studies. In the article it has been considered a brief history of the development of relationships between experimental studies and psychometrics, from the end of 19th century to the present. The historical view allows understanding problems in the use of experimental tasks for assessing individual differences and obstacles to the widespread of use psychometric models in experimental studies. Several recommendations are proposed to improve the accuracy of measurements of individual differences in cognitive abilities and processes, from psychometric perspectives.

Keywords psychometrics, cognitive psychology, experimental psychology, reliability, individual differences, analysis of reaction time, inhibitory function

For citing Kuzmina Yu.V. (2023) Psikhometrika i kognitivnye issledovaniya: protivorechiya i vozmozhnosti kooperatsii [Psychometrics and Cognitive Research: Contradictions and Possibility for Cooperation]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 113–144. https://doi.org/10.17323/vo-2023-16875

Психометрика как наука и область деятельности, связанная с измерением способностей, психологических конструктов и черт, с момента своего зарождения была тесно связана с психологией [Galton, 1879; Cattell, Galton, 1890; Borsboom, 2006], но это не значит, что психометрические подходы, в особенности то, что называется современной теорией тестирования (*Item Response Theory*, IRT), легко усваивались различными направлениями психологии [Borsboom, 2006]. Меньше других, в частности меньше социальной психологии и психологии личности, использует психометрические подходы и модели современной теории тестирования когнитивная психология.

Рассматривая специфику взаимоотношений между когнитивной психологией и психометрикой, следует иметь в виду изменения в содержании понятия «психометрика». На первых этапах развития психометрика представляла собой деятельность по измерению способностей — в первую очередь интеллекта, а также образовательных достижений и индивидуальных различий в способностях, что подразумевало прежде всего разработку и применение стандартизированных тестов. В настоящее время под психометрикой, скорее, подразумевается деятельность по разработке и применению статистических моделей, которые могут использоваться в образовательных или психологических исследованиях для моделирования взаимосвязи между латентными переменными и наблюдаемым поведением. Чаще других применяются модели современной теории тестирования или моделирование структурными уравнениями (*Structural Equation Modeling*, SEM) [Wijsen, Borsboom, Alexandrova, 2022].

Двум разным этапам развития психометрики соответствуют специфические проблемы в ее взаимоотношениях с когнитивной психологией. На первых этапах развития имели место трудности во взаимодействии исследований индивидуальных различий с экспериментальными исследованиями. В настоящее время проблемы возникают из-за использования неподходящих статистических моделей при измерении когнитивных конструктов, например из-за применения в когнитивных исследованиях устаревших подходов классической теории тестирования вместо современной теории тестирования.

Основным препятствием к тесному сотрудничеству между психометрикой и когнитивной психологией является их принадлежность к разным исследовательским традициям, или исследовательским парадигмам, в психологии: когнитивные исследования чаще проводятся в русле экспериментальной парадигмы, а психометрика как средство для оценки индивидуальных различий в интеллекте, в образовательных достижениях и т.п. развивалась в парадигме корреляционных исследований [Borsboom et al., 2009; Cronbach, 1957].

Цель исследований, проводимых в рамках экспериментальной парадигмы, обычно состоит в выделении общих закономерностей или эффектов, для чего могут сравниваться те или иные показатели в контрольной и экспериментальной группах (например, время, потраченное на выполнение заданий, т.е. время ответа) или в двух или нескольких экспериментальных условиях [Smedt de, Gilmore, 2010; Corneille, Mierop, Unkelbach, 2020]. При этом индивидуальные различия между участниками внутри групп считаются «шумом», который по возможности надо свести к минимуму [Borsboom et al., 2009]. Надежность эксперимента определяется тем, насколько он поддается воспроизведению. Чем меньше различий между участниками, чем больше среди участников тех, кто демонстрирует искомый эффект, тем более надежен полученный эффект и, соответственно, эксперимент. В этой парадигме люди взаимозаменяемы, поскольку предполагается, что выделяемые закономерности и процессы одинаковы для всех людей — по крайней мере для всех людей, не имеющих отклонений.

В рамках парадигмы корреляционных исследований и исследований индивидуальных различий, наоборот, гомогенность исследуемой выборки и низкий уровень межиндивидуальной дисперсии являются показателем неуспеха. Чем сильнее люди различаются между собой, чем выше уровень межиндивидуальной дисперсии, тем лучше. Задача состоит в том, чтобы как можно надежнее оценить уровень межиндивидуальных различий. Надежность теста или инструмента в таком случае понимается как способность одинаковым образом

ранжировать участников или как способность измерить оцениваемую способность с минимальной ошибкой [Dunn, Baguley, Brunsden, 2014; Cronbach, Shavelson, 2004]. Таким образом, специфика экспериментальных и корреляционных исследований определяет ограничения в кооперации между ними.

Взаимоотношения исследований индивидуальных различий и экспериментальных исследований станут понятнее, если рассмотреть их в исторической перспективе. Такое рассмотрение покажет, во-первых, что разрыв между когнитивными исследованиями и психометрикой возник не сразу, но связан с разной исследовательской логикой в двух подходах, и во-вторых, что этот разрыв не предопределен — а следовательно, может быть преодолен. На всем протяжении развития психологии многие исследователи подчеркивали возможность сближения рассматриваемых подходов, которое может обогатить психологию в целом.

1. Всегда ли экспериментальные исследования и исследования индивидуальных различий были разделены

Родоначальником исследований индивидуальных различий и первым психометриком считается Ф. Гальтон [Goldstein, 2012; Ludlow, 1998]. Он, в частности, предположил, что индивидуальные различия в сенсомоторных реакциях являются проявлением индивидуальных различий во врожденных способностях [Galton, 1883]. Он разработал систему тестирования некоторых сенсомоторных данных и организовал работу антропометрических лабораторий, которые за несколько лет собрали данные измерений около 17 тыс. человек. Уже в XX в. исследователи проанализировали данные, собранные Ф. Гальтоном, и нашли некоторые из них достаточно надежными для использования [Johnson et al., 1985].

Идеями измерения индивидуальных различий вслед за Ф. Гальтоном воодушевился известный американский психолог, одно время обучавшийся у В. Вундта, Джеймс Маккин Кеттэлл. Он считал, что психология должна развиваться и как наука экспериментальная (к этой традиции он относил исследования Вундта по измерению времени реакции и психофизиков с их исследованиями связи силы стимулов и ощущений), и как наука, использующая тесты и измерения: «Психология не может достичь достоверности и точности физических наук, если она не опирается на эксперимент и измерения. Шаг в этом направлении можно было бы сделать, применив ряд ментальных тестов и измерений к большому числу людей. Результаты будут иметь значительную научную ценность для открытия постоянства психических процессов, их взаимозависимости и их изменчивости при разных обстоятельствах»[1].

[1] "Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this direction could be made by applying a series of mental tests and mea-

Дж. Кеттэлл разработал широкомасштабную программу тестирования способностей, которая должна была быть внедрена в американских колледжах. Программа предполагала сбор результатов о выполнении студентами многочисленных заданий, например определения скорости движений, минимально различаемой разницы в весе, выполнения задания на называние цветов и др. [Cattell, Galton, 1890]. Дж. Кеттэлл подчеркивал, что каждый участник должен пройти достаточно много заданий каждого типа, после чего должны быть рассчитаны средние показатели, дисперсия, максимум и минимум, все эти параметры должны описывать способности каждого тестируемого и предсказывать его академические успехи. Результаты осуществления этой программы не оправдали ожиданий: показатели выполнения студентами предложенных тестов не коррелировали ни друг с другом, ни с академическими успехами испытуемых [Wissler, 1901].

Тем не менее в начале становления и развития психологии экспериментальные исследования и исследования индивидуальных различий скорее дополняли друг друга, чем противопоставлялись. Например, перед экспериментатором выдвигалась задача выявить источник обнаруженных в эксперименте индивидуальных различий в результативности испытуемых: являются ли они следствием различий во врожденных способностях или эффектом обучения и практики [Wells, 1912]. Э. Боринг считал тесты сокращенной версией психологических экспериментов, а Г. Годдард подчеркивал, что разницу между тестами и экспериментами определяет в основном способ использования их результатов [Terman, 1924].

Однако по мере развития теорий и методологических подходов и накопления эмпирических данных различия между экспериментальными исследованиями и исследованиями индивидуальных различий углублялись, и в 1920-х годах научное сообщество уже ясно осознавало специфику подходов и результатов в экспериментальных исследованиях и в тестах. Л. Термен описывал следующие различия: тесты имеют дело с оценкой индивидуальных различий, а не с выявлением общих законов; тесты применяются к большому числу субъектов и нацелены на быстрое определение состояния и поведения, а не внутреннего психологического содержания; результаты тестов, хотя и имеют научную ценность, как правило, менее точны, чем результаты психологических экспериментов [Ibid.].

surements to a large number of individuals. The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances" [Cattell, Galton, 1890. P. 373].

Вряд ли можно выделить какую-то одну причину появления, а затем и углубления разрыва между экспериментальными исследованиями и применением тестов. Показательна история их отношений в США: там с середины 1910-х годов тесты активно внедрялись во все сферы деятельности, их широко применяли в образовании, армии, бизнесе, рекламе. Популярность тестов была обусловлена многими политическими и экономическими обстоятельствами [Sokal, 1987]. С одной стороны, распространение тестирования позволило большому числу психологов найти себе достойную работу и зарплату, а психологии — доказать свою общественную значимость [Шульц, Шульц, 1998]. Некоторые исследователи считают, что тестирование в США стало технологией, так что американские психологи могли позиционировать свою деятельность в качестве такой же полезной, как инженерное дело, и такой же уважаемой, как медицина [Brown, 1992]. С другой стороны, с массовым внедрением тестирования деятельность психометриков приобретала характер, существенно отличающийся от чисто академических исследований.

Распространившийся в те же годы в США бихевиоризм провозгласил, что психология должна стать одной из ветвей естественных наук, а для этого она должна заниматься изучением только наблюдаемого поведения, используя экспериментальные методы [Watson, 1913]. Психология должна была строиться по образцу естественных наук, поэтому основным исследовательским методом должен был стать эксперимент. Бихевиоризм не интересуется индивидуальными различиями, для него есть только общие законы поведения, универсальные для человека и для животных [Braat et al., 2020]. Экспериментальная психология и психометрика продолжали активно развиваться, но отдельно друг от друга.

В середине 1950-х годов в психологии окончательно оформилось представление о психологах-экспериментаторах и исследователях индивидуальных различий (психометриках) как о двух разных классах исследователей, отношения между которыми характеризуются соперничеством и непониманием: «Если психометрик и экспериментатор в чем-то и согласны — а в этом есть определенное сомнение, — то только в одном: каждый из них считает, что другой играет в низшей лиге»[2].

Чуть позже, в 1957 г., вышла известная статья Л. Кронбаха, в которой он констатировал разделение психологии на две ветви: корреляционную психологию и экспериментальную психологию [Cronbach, 1957]. Л. Кронбах подчеркивал, что корреляционные психологи, к которым он относит, например,

[2] "If the psychometric researcher and the experimentalist agree on anything, and there is some doubt about this, it is that the other kind of psychologist plays in the other league (class B)" [Bindra, Scheier, 1954. P. 69].

исследователей психологии развития, индивидуальных различий, личности, хотя и занимаются исследованиями в разных областях психологии, получают схожую подготовку и образование. А экспериментальные психологи получают иную подготовку и могут заниматься экспериментальными исследованиями, не имея знаний в теории тестирования или дифференциальной психологии. О взаимоотношениях представителей двух ветвей психологии Л. Кронбах писал: «Психологи, занимающиеся исследованиями личности, детской и социальной психологии, пошли одним путем, исследователи восприятия и обучения пошли другим, и страна между ними превратилась в пустыню»[3]. Л. Кронбах полагал, что будущее психологии связано с объединением этих двух ветвей, поскольку они могут обогатить друг друга. В частности, он писал, что объединенная психология должна исследовать и различия между субъектами, и различия между условиями, и взаимодействие между субъектом и ситуацией. Исследователи, работающие в корреляционной парадигме, могут помочь исследователям-экспериментаторам выработать новое понимание межиндивидуальных различий, а также предложить новые методы, такие как факторный анализ.

Оживлению дискуссии о различиях в двух исследовательских парадигмах и о необходимости их объединения в 1950-е годы могли способствовать несколько факторов. Первый фактор — это, несомненно, изменения, произошедшие в психологии. В начале 1950-х годов в психологии началось движение, названное в исследовательской литературе когнитивной революцией, что означало возвращение в психологию исследований познавательных процессов [Miller, 2003]. Термин «когнитивная революция» получил широкое распространение, но некоторые авторы считают, что речь шла, скорее, не о революции, а о развитии идей необихевиоризма и отказе от радикального бихевиоризма [Watrin, Darwich, 2012; Moore, 1999]. Из всех направлений психологии именно когнитивная психология (этот термин закрепился уже в 1960-е годы), по мнению некоторых исследователей, теснее всего связана с идеями бихевиоризма и, по сути, является эволюцией необихевиоризма, по крайней мере с точки зрения методологии [Moore, 1996; Watrin, Darwich, 2012]. Отказ от радикального бихевиоризма означал оживление интереса к исследованиям способностей и индивидуальных различий, что отразилось на всех областях психологии [Lamiell, 1992; Royer, 2006].

[3] "The personality, social, and child psychologists went one way; the perception and learning psychologists went the other; and the country between them turned into desert" [Cronbach, 1957. P. 673].

С другой стороны, большие изменения происходят и в психометрике. Когнитивная революция в психологии происходила практически одновременно с рождением современной теории тестирования. Ее основу составили работы, которые в 40–50-х годах XX в. вели параллельно разные исследователи: Ф. Лорд и А. Бирнбаум в США, Г. Раш в Дании, П. Лазерфельд в Австрии [Lord, 1953; Rasch, 1960; Lazerfeld, 1950; Birnbaum, 1958]. В некотором роде разработка моделей современной теории тестирования — это поиск объективности в измерении, попытка преодолеть ограничения классической теории тестирования и создать статистические модели, которые бы удовлетворяли условиям объективного измерения, где баллы оценки латентных способностей не зависят от набора заданий и выборки [Rasch, 1968]. Казалось бы, возвращение когнитивных процессов в психологию и развитие теории объективных измерений в психометрике могут способствовать объединению двух подходов, но оно не произошло.

Нельзя сказать, что таких попыток не было. Необходимость и перспективы сотрудничества когнитивной психологии и психометрики периодически обсуждались и продолжают обсуждаться [Sternberg, 1981; Glaser, 1981; Embretson, 1994; Embretson, Gorin, 2001]. Но складывается впечатление, что в таком объединении долгое время были больше заинтересованы психометрики, чем когнитивные психологи: прежде всего обсуждалось, как когнитивная психология может помочь разработчикам тестов и усилить сферу тестирования [Embretson, 1994]. Для этого разработка заданий тестов должна происходить с опорой на теории когнитивных процессов [Embretson, Gorin, 2001].

Позже в психометрии усилилось влияние идей современной теории тестирования и применения более продвинутых статистических моделей для анализа результатов когнитивных тестов [Rouder, Haaf, 2019; Maas van der et al., 2011; Rouder, Kumar, Haaf, 2019]. Но, по мнению некоторых психометриков, проникновению психометрики в психологию препятствовало отсутствие в психологии сильных теорий [Borsboom, 2006]. Имеются в виду теории, требующие применения какой-то конкретной психометрической модели. Такое предположение касалось прежде всего исследований в психологии личности, но его можно отнести и к когнитивным исследованиям.

Интерес когнитивных психологов к исследованию индивидуальных различий в когнитивных процессах оживился с появлением возможности анализировать нейрофизиологические корреляты тех или иных процессов [Gevins, Smith, 2000; Drew, Vogel, 2008]. Кроме того, в когнитивных исследованиях стали активно использоваться техники конфирматорного факторного анализа и моделирования структурными уравнениями для

исследования факторной структуры сложных когнитивных конструктов [Kane et al., 2004; Friedman, Miyake, 2017].

Таким образом, мы можем наблюдать сближение когнитивной психологии — в той ее части, которая отходит от методологии необихевиоризма, — и психометрики. При этом существующие подходы к оценке индивидуальных различий в когнитивных исследованиях подвергаются критике с учетом уже сложившихся традиций и практик [Hedge, Powell, Sumner, 2018; Goodhew, Edwards, 2019].

2. Сложившиеся традиции и практики анализа результатов исследования

Помимо уже указанных выше особенностей понимания надежности и отношения к межиндивидуальной дисперсии, между когнитивными исследованиями, проводимыми в экспериментальной парадигме, с одной стороны, и исследованиями индивидуальных различий и психометрикой — с другой, существуют и другие различия.

2.1. Трудность заданий: фиксированная или вариативная

Когнитивные исследования и психометрика по-разному подходят к конструированию заданий. В частности, в когнитивных исследованиях часто используются так называемые элементарные когнитивные задания — задачи с низкой трудностью, разработанные в рамках информационно-процессуального подхода [Simon, 1979]. Предполагается, что любой человек при отсутствии ограничений по времени может выполнить такое задание, поэтому, как правило, точность их выполнения в среднем очень высокая и приближается к 100% [Jensen, 2006; Rouder, Haaf, 2019]. Ошибки, совершаемые в таких заданиях, ничего не говорят ни об особенностях процесса решения, ни о способности респондента, а могут интерпретироваться как «шум». В ситуации, когда точность выполнения задания приближается к 100%, информация о правильности выполнения задания не может быть использована для оценки способности респондента или различий между группами или условиями. Поэтому в качестве основного индикатора выполнения заданий чаще всего используется время ответа (или время реакции) [Whelan, 2008; Baayen, Milin, 2010].

В психометрических исследованиях, наоборот, в тесты включаются задания разной трудности и способность респондента анализируется с учетом точности (правильности) выполнения задания (например, при оценке образовательных достижений [Ackerman, Gierl, Walker, 2003]) или ответов в тестах психологических черт с использованием шкал Ликерта [Spens, Owens, Goodyer, 2012]. Одно из преимуществ психометрических моделей в рамках современной теории тестирования состоит

в моделировании вероятности правильного выполнения задания с учетом и способности респондента, и трудности задания [Hambleton, 1989].

Описанные различия в конструировании тестов, безусловно, не являются абсолютными и непреодолимыми. В когнитивных исследованиях, как и в психометрике, могут использоваться задания с разным уровнем трудности, в которых возможна градация точности ответа [Dietrich, Huber, Nuerk, 2015]. В то же время в психометрических исследованиях для более точной оценки способности респондентов все чаще используют информацию о времени ответа, хотя, как правило, совместно с информацией о точности [Linden van der, 2009; Molenaar et al., 2015].

2.2. Анализ времени ответов: процессные модели или моделирование латентных конструктов

Время выполнения задания часто используется в когнитивных исследованиях для анализа результатов выполнения заданий и описания характеристик когнитивных процессов. При этом учитывается, что общее время ответа не равно скорости оцениваемых процессов. Еще в 1890 г. Дж. Кеттэлл указывал на трудности при анализе времени ответа, поскольку общее время ответа не дает информации о скорости отдельных процессов, протекающих во время выполнения задания. Для того чтобы отделить друг от друга разные процессы и оценить скорость протекания отдельного процесса, используются различные математические модели, которые условно можно назвать процессными, поскольку их задача — отделить целевой процесс от сопутствующих. К этому классу можно отнести диффузионную модель [Ratcliff, Smith, McKoon, 2015; Воронин и др., 2020].

Несмотря на существование процессных моделей, результаты тестирования в когнитивных исследованиях до сих пор чаще всего анализируются с применением оценки среднего времени ответа (или медианы) для определенных условий и групп и сравнения этих показателей. Иногда используется время ответа только для правильно выполненных заданий, может применяться трансформация времени ответа, например логарифмирование [Whelan, 2008; Baayen, Milin, 2010; Lo, Andrews, 2015]. При таком подходе время ответа отождествляется со скоростью процессов и становится возможным индикатором способности респондента — при условии одинаковой трудности заданий [Dodonova, Dodonov, 2013].

В рамках современной теории тестирования анализ времени ответа не получил широкого распространения. Однако в последние годы психометрики также стали предлагать модели для анализа времени ответа. При этом, в отличие от процессных моделей, используемые в психометрике модели со време-

нем ответа используются для более точной оценки способности респондента. Так как для IRT критически важна отдельная оценка параметров респондента и задания, рассматриваемые модели выделяют отдельные параметры для оценки трудности задания и его временной нагрузки, способности респондента и его быстроты [Goldhammer et al., 2014; Molenaar et al., 2015; Bolsinova, Tijmstra, 2018; Maas van der et al., 2011].

В настоящее время существует много разновидностей психометрических моделей с использованием времени ответа и точности, в некоторых из них эти параметры могут моделироваться вместе, но могут быть не связаны друг с другом [De Boeck, Jeon, 2019]. Таковы, например, обобщенные линейные иерархические модели [Linden van der, 2007]. В них применяются три иерархических уровня моделирования: внутрииндивидуальный, индивидуальный и уровень популяции (межиндивидуальный). Для точности и времени ответа создаются две разные модели, на каждом из уровней выделяются параметры заданий и параметры респондента.

В других моделях моделируется связь между двумя латентными переменными для точности и скорости: например, в *Bivariate Generalized IRT model* (B-GLIRT-модель) [Molenaar, Tuerlinckx, van der Maas, 2015] или в модели локальной зависимости [Bolsinova, Tijmstra, 2018]. В B-GLIRT-модели идентифицируются два латентных фактора для точности и скорости, а также моделируется функция связи между ними [Molenaar, Tuerlinckx, van der Maas, 2015].

Наконец, есть психометрические модели, в которых время ответа является независимой переменной для точности [De Boeck, Jeon, 2019]. Таковы, в частности, модели со смешанными эффектами для точности как зависимой переменной. С их помощью было показано, что связь между временем ответа и вероятностью дать правильный ответ зависит от типа задач и уровня подготовленности респондента [Goldhammer et al., 2014].

Некоторые из предлагаемых психометриками моделей представляют собой доработанные варианты диффузионной процессной модели (например, *positive ability model* [Maas van der et al., 2011]). Диффузионная модель является разновидностью двухпараметрической IRT-модели и при определенных условиях может быть использована для оценки способностей [Tuerlinckx, De Boeck, 2005].

Итак, в психометрике время ответа позволяет уточнить или выделить дополнительные параметры заданий и используется для оценки способности респондента, а в когнитивной психологии время ответа служит для описания когнитивного процесса, стоящего за выполнением заданий. В некоторых случаях в когнитивной психологии время ответа также может быть ис-

пользовано для описания способности респондента, но практически всегда это происходит без учета трудности задания.

2.3. Количество заданий: фиксированное или меняющееся

Определяя количество заданий, достаточное для тестирования той или иной способности или черты или для оценки определенного процесса, психометрики и исследователи когнитивных процессов исходят из разных целей. Стандартизированные тесты и психологические опросники, как правило, содержат фиксированное количество заданий, которое не меняется от исследования к исследованию. При этом для идентификации латентного конструкта, т.е. измеряемой способности, часто используется небольшое количество заданий. Например, для идентификации латентного конструкта с помощью конфирматорного факторного анализа достаточно трех заданий. Чем меньше заданий, тем быстрее испытуемый пройдет тест. Время, необходимое для тестирования, оказывается принципиальным фактором при сборе данных на больших выборках — а в корреляционных исследованиях размер выборки важен. Для экзаменов же с высокими ставками, конечно, необходимо больше заданий, чтобы свести к минимуму ошибку при оценке способности экзаменуемого. В любом случае в стандартизированных тестах количество заданий не меняется. Более того, психометрики настаивают на том, что изменение количества заданий требует новой идентификации модели и оценки психометрических свойств инструмента [Rouder, Haaf, 2019; Kleka, Soroko, 2018].

Так как в когнитивных исследованиях стоит задача элиминировать «шум», а размер выборки долгое время был не очень важен, здесь, как правило, для оценки того или иного процесса или функции используется много заданий. При этом количество заданий может существенно варьировать от исследования к исследованию. Например, одно из самых популярных заданий для исследования процесса подавления действия нежелательных стимулов (ингибиторной функции) и когнитивного контроля — вербально-цветовой тест Струпа [Stroop, 1935]. В классическом исследовании Д. Струпа предлагалось по 100 заданий для каждого условия: в первом условии слово, обозначающее цвет, не совпадает с цветом шрифта, во втором — слово, обозначающее цвет, напечатано черным шрифтом. Затем тест стали использовать другие исследователи, они меняли число заданий и другие параметры эксперимента, например вводили условие, при котором значение слова, обозначающего цвет, и цвет шрифта совпадают. В разных исследованиях может быть от 10 до 100 заданий [Scarpina, Tagini, 2017].

2.4. Измерительная инвариантность: проверять или нет

Измерительная инвариантность, т.е. одинаковая работа заданий на разных выборках, — очень важный принцип психометрики [Putnick, Bornstein, 2016; Leitgöb et al., 2023]. Чтобы сравнивать группы по уровню способности или степени выраженности какого-то латентного конструкта, необходимо убедиться, что используемый инструмент (задания) работает одинаково во всех рассматриваемых группах. Измерительная инвариантность может быть проверена с помощью конфирматорного факторного анализа или в рамках современной теории тестирования [Kim, Yoon, 2011; Meade, Lautenschlager, 2004].

Чтобы убедиться в измерительной инвариантности используемого инструмента, необходимо проверить три предположения: о сохранении факторной структуры в группах (конфигуральная инвариантность), об одинаковых факторных нагрузках для сравниваемых групп (метрическая инвариантность) и об одинаковых интерцептах для метрических индикаторов или пороговых значений для категориальных индикаторов (скалярная инвариантность).

В рамках современной теории тестирования контроль измерительной инвариантности равнозначен проверке DIF (*Different Item Functioning*). Проверить DIF — значит оценить инвариантность функции ответа на задания: будет ли одна и та же модель подходить всем группам респондентов с одними и теми же параметрами заданий [Kim, Yoon, 2011]. Проверка измерительной инвариантности, или оценка DIF, — распространенная практика оценки психометрических свойств, без доказанной инвариантности сравнивать группы нельзя [Putnick, Bornstein, 2016].

Сравнение групп участников по успешности выполнения заданий нередко выполняется и в когнитивных исследованиях [Passolunghi, Siegel, 2001], но измерительная инвариантность инструментария контролируется обычно только на клинических выборках. Д. Борсбум, рассматривая вопросы взаимодействия психометрики и психологии, указывал на проблему измерительной инвариантности тестов интеллекта [Borsboom, 2006]. Впрочем, с тех пор появилось немало исследований измерительной инвариантности когнитивных тестов: тестов интеллекта, тестов для оценки рабочей памяти [Wicherts, 2016; Willoughby et al., 2012]. При этом проверка измерительной инвариантности производится преимущественно в исследованиях интеллекта — конструкта, который относится скорее к психометрике, чем собственно к когнитивной психологии.

Итак, когнитивная психология начинает более активно заниматься исследованиями индивидуальных различий, но между ней и психометрикой сохраняется определенный разрыв, обусловленный различиями в исследовательских подходах и

сложившимися традициями. Вместе с ростом интереса исследователей когнитивных процессов к оценке индивидуальных различий в тестируемых способностях увеличивается и число публикаций, в которых обсуждаются проблемы, связанные с использованием для этой цели привычных для когнитивных психологов инструментов, например теста Струпа или теста флангов.

3. Проблемы измерений индивидуальных различий в когнитивных исследованиях с точки зрения психометрики

3.1. Ограничение трудности заданий

Ограничение трудности заданий может иметь несколько неблагоприятных последствий, связанных друг с другом. Во-первых, в результате отбора заданий с низким уровнем трудности сокращается дисперсия точности выполнения заданий на межиндивидуальном уровне. Из-за этого может снижаться надежность используемых тестов и уменьшаться возможная корреляция между измеряемым конструктом и другими переменными. Во-вторых, вследствие ограничения трудности заданий возникает необходимость анализировать время ответа, что создает дополнительные методологические трудности. Далее мы рассмотрим кратко обе проблемы.

3.2. Низкая надежность используемых инструментов

Многие инструменты, применяемые в когнитивных исследованиях, не подходят для оценки индивидуальных различий, поскольку имеют низкую надежность как на гомогенной выборке, так и на гетерогенной [Hedge, Powell, Sumner, 2018; Pronk et al., 2023]. Особенно много критики звучит в адрес заданий, используемых для оценки функции подавления [Rey-Mermet, Gade, Oberauer, 2018; Rouder, Kumar, Haaf, 2019]. Некоторые исследователи считают, что низкая надежность инструментов может быть связана с особенностями расчета индивидуальных баллов во многих заданиях на измерение функции подавления [Hedge, Powell, Sumner, 2018].

Обычно при расчете индивидуальных показателей по тестам для оценки функции подавления используют разницу в средних показателях времени ответа между конгруэнтными и неконгруэнтными заданиями или между нейтральными и неконгруэнтными. Чем больше разница между конгруэнтными и неконгруэнтными заданиями, тем ниже уровень ингибиторной функции. Отмечалось, что в целом надежность баллов, рассчитанных как разница времени ответа между двумя условиями, всегда ниже, чем надежность баллов в каждом отдельном условии, — возможно, из-за того, что дисперсия разницы баллов ниже, чем дисперсия в каждом из условий, особенно если корреляция показателей высокая [Caruso, 2004; Eide et al., 2002; Edwards, 2001].

Низкая надежность и недостаточный уровень дисперсии в некоторых тестах могут также приводить к снижению или отсутствию связи между результатами тестов, измеряющих, по идее, один и тот же конструкт. Например, показано, что результаты вербально-цветового теста Струпа имеют очень низкую корреляцию с показателями флангового теста, хотя оба инструмента измеряют устойчивость к нерелевантным стимулам или подавление доминирующего стимула [Rey-Mermet, Gade, Oberauer, 2018; Stahl et al., 2014]. Возможны два объяснения низкой корреляции показателей. Первое: эти два типа заданий действительно оценивают разные способности, и поэтому показатели не коррелируют друг с другом. Второе: низкая корреляция связана с низкой надежностью вследствие большой ошибки измерения и/или с низким уровнем дисперсии для одной или двух переменных. При этом степень снижения корреляции зависит от числа заданий и вариабельности между ними [Rouder, Haaf, 2019].

3.3. Проблемы с анализом времени ответа

Применительно ко времени ответа наиболее часто обсуждаются следующие методологические вопросы: время ответа обычно не имеет нормального распределения, эта переменная не имеет отрицательных значений, и ее распределение имеет правый скос [Whelan, 2008]. Кроме того, часто при анализе времени ответа обнаруживаются выбросы, которые могут исказить оценку средних эффектов [Heathcote, Popiel, Mewhort, 1991; Rousselet, Wilcox, 2020; Baayen, Milin, 2010]. В итоге используемые средние показатели времени ответа могут не отражать реальной тенденции и искажать оценку эффектов [Speelman, McGann, 2013].

В одном из исследований с использованием теста Струпа были проанализированы средние показатели времени ответа по каждому из условий. Выявлены значимые различия по времени ответа между неконгруэнтными и нейтральными заданиями (эффект интерференции), но не между конгруэнтными и нейтральными (эффект фасилитации отсутствует). Оказалось, что время ответа соответствует, скорее, экспоненциально модифицированному распределению Гаусса (*ex-Gaussian*). С учетом характера распределения были рассчитаны иные показатели мер средней тенденции и получены иные результаты в отношении эффектов, при этом подтверждены оба эффекта — и фасилитации, и интерференции [Heathcote, Popiel, Mewhort, 1991].

Для того чтобы решить проблему с отклонениями от нормального распределения, некоторые исследователи рекомендуют отказываться от использования средних показателей и параметрических методов анализа, а рассчитывать, напри-

мер, медиану вместо среднего [Whelan, 2008; Speelman, McGann, 2013]. Однако использование медианы тоже не всегда «спасает». В частности, показано, что на маленькой выборке медианные оценки могут быть более смещенными, чем среднее. Кроме того, медиану не рекомендуют использовать, например, если в исследовании сравниваются условия, тестируемые с помощью разного количества заданий [Rousselet, Wilcox, 2020].

Еще один распространенный вариант борьбы с отклонениями от нормального распределения при использовании времени ответа — трансформация переменной. Наиболее часто применяют логарифмирование [Schramm, Rouder, 2019; Lo, Andrews, 2015]. Трансформация позволяет применять параметрические методы анализа, к которым привыкли большинство психологов. Кроме того, она может быть полезна для обнаружения небольших эффектов [Schramm, Rouder, 2019]. Однако следует учитывать, во-первых, что происходящее после трансформации изменение шкалы измерения времени не всегда имеет смысл с точки зрения теории и интерпретации полученных результатов [Lo, Andrews, 2015]. Во-вторых, оценки эффектов на «сырой» шкале и на трансформированной шкале могут различаться, например на логарифмированной шкале могут обнаружиться значимые эффекты, которых нет на «сырой» шкале. Некоторые исследователи также отмечают, что иногда разные виды трансформации используются для *p-hacking* — для манипуляции данными с целью получить значимые эффекты [Moris Fernández, Vadillo, 2020].

Если цель исследования заключается в оценке различий между условиями, нет необходимости стремиться к нормальному распределению (при этом необходимо принять решение о том, что делать с выбросами). Однако если цель состоит в оценке связей между временем ответа и другими переменными, использование трансформации оправданно [Schramm, Rouder, 2019].

Оценка индивидуальных различий только на основании времени ответа может быть проблематичной еще и потому, что надежность разных показателей с использованием времени ответа ниже надежности индикаторов с использованием точности [Draheim et al., 2019; Dietrich et al., 2016; Saville et al., 2011].

3.4 Изменение количества заданий

Варьирование числа заданий может существенно изменить показатели корреляций, надежности и размеров эффекта в случае использования подходов классической теории тестирования вследствие нарушения допущения портативности (*portability*) — неизменности значений, полученных с помощью инструмента, в популяции независимо от размера выборки [Rouder, Haaf,

2019]. Так, показано, что увеличение количества заданий для каждого из условий в вербально-цветовом тесте Струпа ведет к нарастанию размеров эффекта и повышению надежности — этот эффект известен в классической теории тестирования. Увеличение длины шкалы приводит к уменьшению ошибки измерения [Rouder, Haaf, 2019]. Следовательно, для того чтобы сравнивать размеры эффекта, полученные в разных исследованиях, необходимо учитывать разницу в числе заданий, а не только размер выборки.

3.5. Использование агрегированных данных

В большинстве когнитивных исследований — как в экспериментальных, так и при анализе индивидуальных различий — используются агрегированные показатели времени ответа (например, среднее время ответа для респондента или группы респондентов, среднее время для условий) или точности (сумма или пропорция правильных ответов). В частности, для теста Струпа рассчитываются пропорция правильных ответов (или пропорция ошибок) и/или среднее время правильного ответа для каждого условия и средняя разница между конгруэнтными, неконгруэнтными и нейтральными условиями [Schmidt, Besner, 2008; Schichel, Tzeglov, 2018]. Далее эти показатели также могут агрегироваться на уровень выборки или группы [Hepp et al., 1996].

Основная проблема, связанная с использованием агрегированных данных, состоит в потере информации о внутрииндивидуальной вариабельности, т.е. дисперсии времени ответа или точности между заданиями, в то время как такая дисперсия может быть выше межиндивидуальной [Rouder, Haaf, 2019]. В психологии уже давно известны ограничения использования агрегированных данных и возможные негативные эффекты их применения, в частности парадокс Симпсона: связи между двумя агрегированными переменными могут быть прямо противоположными связям между переменными, оцененными не на агрегированном уровне [Kievit et al., 2013]. В когнитивных исследованиях, например, связь между временем ответа и точностью на уровне индивида может отличаться от связи между этими переменными на уровне выборки [Moleenar, Tuerlinckx, van der Maas, 2015]. Для того чтобы учитывать внутрииндивидуальную дисперсию и различия в характере связи между переменными на уровне индивида и на уровне выборки, необходимо применять анализ на уровне заданий с использованием моделей со смешанными эффектами или конфирматорного факторного анализа [Moleenar, Tuerlinckx, van der Maas, 2015; Brauer, Curtin, 2018; Cunnings, 2012]. Однако в когнитивных исследованиях применение таких моделей все еще скорее исключение.

4. Есть ли возможности для взаимодействия

Несмотря на наличие проблем, связанных с измерениями, в когнитивных исследованиях, некоторые исследователи считают, что когнитивная психология и психометрика могли бы наладить более тесное сотрудничество и были бы полезны друг другу. Психометрики могут использовать возможности когнитивной психологии для выдвижения теорий и моделей когнитивных процессов, операционализации конструктов, генерации заданий для тестов [Embretson, Gorin, 2001]. Для когнитивных исследований психометрика может быть источником статистических моделей и подходов для оценки психометрических свойств тестов, а также для анализа полученных результатов с учетом специфики заданий и способности респондентов [Maas van der et al., 2011; Rouder, Haaf, 2019; Heck, Erdfelder, 2016].

Проблема измерений в когнитивной психологии возникает в тот момент, когда экспериментальная парадигма трансформируется в парадигму оценки индивидуальных различий, когда инструменты, обычно работающие в экспериментальных исследованиях, привлекаются для оценки индивидуальных различий. Использование психометрических моделей в когнитивной психологии обсуждается именно применительно к оценке индивидуальных различий, психометрики не претендуют на выявление общих механизмов или законов памяти и восприятия. Основной посыл со стороны психометриков когнитивным психологам можно сформулировать так: если исследователь когнитивных процессов переходит от экспериментов и оценки средних эффектов к тестированию способностей и индивидуальных различий, необходимо пользоваться уже разработанными психометрическими подходами и моделями, чтобы делать это правильно.

Что значит правильно? Психометрики, вероятно, могли бы дать ряд рекомендаций по измерению индивидуальных различий в когнитивных исследованиях.

Во-первых, следует изменить подходы к использованию и разработке инструментов для оценки индивидуальных различий в когнитивных процессах. Не всегда стоит применять методики, работающие в экспериментальных исследованиях, даже если они широко известны и хорошо себя зарекомендовали. Возможно, имеет смысл создавать новые инструменты для исследования индивидуальных различий с учетом ранее выявленных проблем. Например, учитывая сравнительно низкую надежность показателей на основе времени ответа, стоит разрабатывать инструменты, в которых будет изменяться трудность заданий и будет оцениваться точность ответа, а не только время ответа [Draheim et al., 2021]. При использовании популярных экспериментальных методик для оценки индивидуальных различий, возможно, имеет смысл переходить от общих назва-

ний методик к указанию того, что конкретно измеряет данная методика (например, тест Струпа — тест устойчивости к дистракторам) по аналогии со шкалами, применяемыми в исследованиях психологических конструктов. Из имеющихся вариантов экспериментальных методик и условий целесообразно отбирать тот, который имеет наибольшую межиндивидуальную дисперсию и надежность [Goodhew, Edwards, 2019]. Кроме того, для каждой методики необходимо указывать число заданий и учитывать его при интерпретации результатов. Возможно, имеет смысл разрабатывать стандартизированные методики с фиксированным количеством заданий и стимулов. Такие методики есть: например, модифицированная и стандартизированная версия теста Струпа, разработанная в Университете Виктории (*Victoria Stroop test*) [Troyer, Leach, Strauss, 2006], но для анализа результатов используются «классические» методы с агрегированными данными.

Во-вторых, важно при публикации результатов сообщать надежность когнитивных тестов, используемых в исследованиях индивидуальных различий [Parsons, Kruijt, Fox, 2019]. При этом надежность каждого конкретного теста должна быть оценена отдельно, нельзя полагаться на ранее полученные оценки надежности, поскольку они могут зависеть от параметров выборки. Достаточно часто психологи используют в качестве показателя надежности коэффициент альфа Кронбаха, который подразумевает ряд ограничений [Kim, Feldt, 2010; Dunn, Baguley, Brunsden, 2014; Tavakol, Dennick, 2011]. В настоящее время психометрики рекомендуют применять другие показатели надежности, например коэффициент омега [Dunn, Baguley, Brunsden, 2014].

В-третьих, при интерпретации полученных корреляций не стоит полагаться на оценку надежности как на гарантию их точной оценки [Rouder, Kumar, Haaf, 2019]. Высокая надежность используемых шкал может служить такой гарантией только при применении стандартных тестов или шкал, содержащих одни и те же формулировки заданий и одинаковое количество заданий. В экспериментальных методиках такое бывает редко. Поэтому даже при получении высоких показателей надежности нужно учитывать возможность недооценки корреляции на уровне выборки [Ibid.].

В-четвертых, психометрики рекомендуют перестать использовать агрегированные показатели: среднюю точность или среднее время для респондента, или для условия, или для выборки. Их следует заменить моделями, которые учитывают дисперсию между заданиями, например моделями со смешанными эффектами или иерархическими моделями конфирматорного факторного анализа [Rouder, Kumar, Haaf, 2019; Molenaar,

Tuerlinckx, van der Maas, 2015]. В психометрике за последние годы разработано много моделей, учитывающих как точность, так и время ответа, эти модели можно применять для анализа результатов когнитивных тестов. Например, B-GLIRT-модель может учитывать дисперсию как на внутрииндивидуальном уровне, так и на межиндивидуальном. В эту модель могут быть включены параметры ответа и времени ответа для каждого задания, на этой основе можно оценивать латентные способности респондента, его быстроту как латентную характеристику, а также разные типы связи между точностью и скоростью [Molenaar, Tuerlinckx, van der Maas, 2015]. При этом модель допускает возможность включения нелинейных связей между точностью и скоростью и взаимодействие между ними и трудностью задания.

Различия в двух рассмотренных психологических традициях — экспериментальной и дифференциально-измерительной, — возможно, непреодолимы в обозримом будущем не столько из-за разницы в методологических подходах, сколько из-за различий в предмете исследования [Borsboom et al., 2009]. Поэтому, по выражению Д. Борсбума, стоит принять рабочую гипотезу о разделенной психологии.

При этом необходимо учитывать и то общее, что стоит за двумя традициями. Исследователи индивидуальных различий не должны исключать возможность, что некоторые межиндивидуальные различия могут быть порождены системами внутрииндивидуальных процессов, и, наоборот, теории внутрииндивидуальных процессов не исключают возможности межиндивидуальных различий [Borsboom et al., 2009]. Кроме того, важно понимать ограничения каждого подхода. Их наличие означает, что результаты, полученные в экспериментальных исследованиях, не могут и не должны быть приложимы для описания отдельных индивидов. И наоборот, результаты изучения индивидуальных различий не могут быть прямо перенесены на описание внутрииндивидуальных процессов и механизмов [Molenaar, Beltz, 2020].

Благодарности

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

Литература

1. Воронин И.А., Захаров И.М., Табуева А.О., Мерзон Л.А. (2020) Диффузная модель принятия решения: оценка скорости и точности ответов в задачах выбора из двух альтернатив в исследованиях когнитивных процессов и способностей. *Теоретическая и экспериментальная психология*, т. 13, № 2, сс. 6–23.

2. Шульц Д.П., Шульц С.Э. (1998) *История современной психологии*. СПб.: Евразия.
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. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x
4. Baayen R.H., Milin P. (2010) Analyzing Reaction Times. *International Journal of Psychological Research*, vol. 3, no 2, pp. 12–28. https://doi.org/10.21500/20112084.807
5. Bindra D., Scheier I.H. (1954) The Relation between Psychometric and Experimental Research in Psychology. *American Psychologist,* vol. 9, no 2, pp. 69–71. https://doi.org/10.1037/h0062472
6. Birnbaum A. (1958) *On the Estimation of Mental Ability. Series Report no 15, Project no 7755–7723*. Texas: Randolph Air Force Base, TX USAF School of Aviation Medicine.
7. Bolsinova M., Tijmstra J. (2018) Improving Precision of Ability Estimation: Getting More from Response Times. *British Journal of Mathematical and Statistical Psychology*, vol. 71, no 1, pp. 13–38. https://doi.org/10.1111/bmsp.12104
8. Borsboom D. (2006) The Attack of the Psychometricians. *Psychometrika*, vol. 71, no 3, pp. 425–440. https://doi.org/10.1007/s11336-006-1447-6
9. Borsboom D., Kievit R.A., Cervone D., Hood S.B. (2009) The Two Disciplines of Scientific Psychology, or: The Disunity of Psychology as a Working Hypothesis. *Dynamic Process Methodology in the Social and Developmental Sciences* (eds J. Valsiner, P. Molenaar, M. Lyra, N. Chaudhary), New York, NY: Springer, pp. 67–97. https://doi.org/10.1007/978-0-387-95922-1_4
10. Braat M., Engelen J., van Gemert T., Verhaegh S. (2020) The Rise and Fall of Behaviorism: The Narrative and the Numbers. *History of Psychology*, vol. 23, no 3, pp. 252–280. https://doi.org/10.1037/hop0000146
11. Brauer M., Curtin J.J. (2018) Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items. *Psychological Methods*, vol. 23, no 3, pp. 389–411. https://doi.org/10.1037/met0000159
12. Brown J. (1992) *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890–1930*. Princeton, NJ: Princeton University.
13. Caruso J.C. (2004) A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *European Journal of Psychological Assessment*, vol. 20, no 3, pp. 166–171. https://doi.org/10.1027/1015-5759.20.3.166
14. Cattell J.M., Galton F. (1890) Mental Tests and Measurements. *Mind*, vol. os-XV, iss. 59, pp. 373–381. https://doi.org/10.1093/mind/os-XV.59.373
15. Corneille O., Mierop A., Unkelbach C. (2020) Repetition Increases Both the Perceived Truth and Fakeness of Information: An Ecological Account. *Cognition*, vol. 205, December, Article no 104470. https://doi.org/10.1016/j.cognition.2020.104470
16. Cronbach L.J. (1957) The Two Disciplines of Scientific Psychology. *American Psychologist,* vol. 12, no 11, pp. 671–684. https://doi.org/10.1037/h0043943
17. Cronbach L.J., Shavelson R.J. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, vol. 64, no 3, pp. 391–418. https://doi.org/10.1177/0013164404266386
18. Cunnings I. (2012) An Overview of Mixed-Effects Statistical Models for Second Language Researchers. *Second Language Research*, vol. 28, no 3, pp. 369–382. https://doi.org/10.1177/0267658312443651

19. De Boeck P., Jeon M. (2019) An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, vol. 10, February, Article no 102. https://doi.org/10.3389/fpsyg.2019.00102
20. Dietrich J.F., Huber S., Klein E., Willmes K., Pixner S., Moeller K. (2016) A Systematic Investigation of Accuracy and Response Time Based Measures Used to Index ANS Acuity. *PLoS ONE*, vol. 11, no 9, Article no e0163076. https://doi.org/10.1371/journal.pone.0163076
21. Dietrich J.F., Huber S., Nuerk H.-C. (2015) Methodological Aspects to Be Considered When Measuring the Approximate Number System (ANS): A Research Review. *Frontiers in Psychology*, vol. 6, March, Article no 295. https://doi.org/10.3389/fpsyg.2015.00295
22. Dodonova Yu.A., Dodonov Yu.S. (2013) Faster on Easy Items, More Accurate on Difficult Ones: Cognitive Ability and Performance on a Task of Varying Difficulty. *Intelligence*, vol. 41, no 1, pp. 1–10. https://doi.org/10.1016/j.intell.2012.10.003
23. Draheim C., Mashburn C.A., Martin J.D., Engle R.W. (2019) Reaction Time in Differential and Developmental Research: A Review and Commentary on the Problems and Alternatives. *Psychological Bulletin*, vol. 145, no 5, pp. 508–535. https://doi.org/10.1037/bul0000192
24. Draheim C., Tsukahara J.S., Martin J.D., Mashburn C.A., Engle R.W. (2021) A Toolbox Approach to Improving the Measurement of Attention Control. *Journal of Experimental Psychology: General,* vol. 150, no 2, pp. 242–275. https://doi.org/10.1037/xge0000783
25. Drew T., Vogel E.K. (2008) Neural Measures of Individual Differences in Selecting and Tracking Multiple Moving Objects. *The Journal of Neuroscience*, vol. 28, no 16, pp. 4183–4191. https://doi.org/10.1523/JNEUROSCI.0556-08.2008
26. Dunn T.J., Baguley T., Brunsden V. (2014) From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation. *British Journal of Psychology*, vol. 105, no 3, pp. 399–412. https://doi.org/10.1111/bjop.12046
27. Edwards J.R. (2001) Ten Difference Score Myths. *Organizational Research Methods*, vol. 4, no 3, pp. 265–287. https://doi.org/10.1177/109442810143005
28. Eide P., Kemp A., Silberstein R.B., Nathan P.J., Stough C. (2002) Test-Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change. *The Journal of Psychology*, vol. 136, no 5, pp. 514–520. https://doi.org/10.1080/00223980209605547
29. Embretson S. (1994) Applications of Cognitive Design Systems to Test Development. *Cognitive Assessment: A Multidisciplinary Perspective* (ed. C.R. Reynolds), New York, NY: Springer Science+ Business Media, pp. 107–135. https://doi.org/10.1007/978-1-4757-9730-5_6
30. Embretson S., Gorin J. (2001) Improving Construct Validity with Cognitive Psychology Principles. *Journal of Educational Measurement*, vol. 38, no 4, pp. 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x
31. Friedman N.P., Miyake A. (2017) Unity and Diversity of Executive Functions: Individual Differences as a Window on Cognitive Structure. *Cortex*, no 86, pp. 186–204. https://doi.org/10.1016/j.cortex.2016.04.023
32. Galton F. (1879) Psychometric Experiments. *Brain*, vol. 2, no 2, pp. 149–162. https://doi.org/10.1093/brain/2.2.149
33. Galton F. (1883) *Inquiries into Human Faculty and Its Development.* New York, NY: MacMillan. https://doi.org/10.1037/14178-000
34. Gevins A., Smith M.E. (2000) Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. *Cerebral Cortex*, vol. 10, no 9, pp. 829–839. https://doi.org/10.1093/cercor/10.9.829
35. Glaser R. (1981) The Future of Testing: A Research Agenda for Cognitive Psychology and Psychometrics. *American Psychologist*, vol. 36, no 9, pp. 923–936. https://doi.org/10.1037/0003-066X.36.9.923

36. Goldhammer F., Naumann J., Stelter A., Tóth K., Rölke H., Klieme E. (2014) The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights from a Computer-Based Large-Scale Assessment. *Journal of Educational Psychology,* vol. 106, no 3, pp. 608–626. https://doi.org/10.1037/a0034716
37. Goldstein H. (2012) Francis Galton, Measurement, Psychometrics and Social Progress. *Assessment in Education: Principles, Policy & Practice*, vol. 19, no 2, pp. 147–158. https://doi.org/10.1080/0969594X.2011.614220
38. Goodhew S.C., Edwards M. (2019) Translating Experimental Paradigms into Individual-Differences Research: Contributions, Challenges, and Practical Recommendations. *Consciousness and Cognition*, vol. 69, January, pp. 14–25. https://doi.org/10.1016/j.concog.2019.01.008
39. Hambleton R.K. (1989) Principles and Selected Applications of Item Response Theory. *Educational Measurement* (ed. R.L. Linn), New York, NY: Macmillan Publishing Co, Inc; American Council on Education, pp. 147–200.
40. Heathcote A., Popiel S.J., Mewhort D.J. (1991) Analysis of Response Time Distributions: An Example Using the Stroop Task. *Psychological Bulletin,* vol. 109, no 2, pp. 340–347. https://doi.org/10.1037/0033-2909.109.2.340
41. Heck D.W., Erdfelder E. (2016) Extending Multinomial Processing Tree Models to Measure the Relative Speed of Cognitive Processes. *Psychonomic Bulletin & Review*, vol. 23, no 5, pp. 1440–1465. https://doi.org/10.3758/s13423-016-1025-6
42. Hedge C., Powell G., Sumner P. (2018) The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behavior Research Methods*, vol. 50, no 3, pp. 1166–1186. https://doi.org/10.3758/s13428-017-0935-1
43. Hepp H.H., Maier S., Hermle L., Spitzer M. (1996) The Stroop Effect in Schizophrenic Patients. *Schizophrenia Research*, vol. 22, no 3, pp. 187–195. https://doi.org/10.1016/S0920-9964(96)00080-1
44. Jensen A.R. (2006) *Clocking the Mind: Mental Chronometry and Individual Differences*. Amsterdam: Elsevier.
45. Johnson R.C., McClearn G.E., Yuen S., Nagoshi C.T., Ahern F.M., Cole R.E. (1985) Galton's Data a Century Later. *American Psychologist,* vol. 40, no 8, pp. 875–892. https://doi.org/10.1037/0003-066X.40.8.875
46. Kane M.J., Hambrick D.Z., Tuholski S.W., Wilhelm O., Payne T.W., Engle R.W. (2004) The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General,* vol. 133, no 2, pp. 189–217. https://doi.org/10.1037/0096-3445.133.2.189
47. Kievit R.A., Frankenhuis W.E., Waldorp L.J., Borsboom D. (2013) Simpson's Paradox in Psychological Science: A Practical Guide. *Frontiers in Psychology*, vol. 4, August, Article no 513. https://doi.org/10.3389/fpsyg.2013.00513
48. Kim E.S., Yoon M. (2011) Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 18, no 2, pp. 212–228. https://doi.org/10.1080/10705511.2011.557337
49. Kim S., Feldt L.S. (2010) The Estimation of the IRT Reliability Coefficient and Its Lower and Upper Bounds, with Comparisons to CTT Reliability Statistics. *Asia Pacific Education Review*, vol. 11, no 2, pp. 179–188. https://doi.org/10.1007/s12564-009-9062-8
50. Kleka P., Soroko E. (2018) How to Avoid the Sins of Questionnaire Abridgement — Guideline. *PsyArXiv*. https://doi.org/10.31234/osf.io/8jg9u
51. Lamiell J.T. (1992) Personality Psychology and the Second Cognitive Revolution. *American Behavioral Scientist*, vol. 36, no 1, pp. 88–101. https://doi.org/10.1177/0002764292036001008

52. Lazarsfeld P.F. (1950) The Logical and Mathematical Foundation of Latent Structure Analysis. *Studies in Social Psychology in World War II. Vol. IV: Measurement and Prediction* (eds S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld), Princeton: Princeton University, pp. 362–412.
53. Leitgöb H., Seddig D., Asparouhov T., Behr D., Davidov E., de Roover K. et al. (2023) Measurement Invariance in the Social Sciences: Historical Development, Methodological Challenges, State of the Art, and Future Perspectives. *Social Science Research*, vol. 110, January, Article no 102805. https://doi.org/10.1016/j.ssresearch.2022.102805
54. Linden van der W.J. (2009) Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, vol. 46, no 3, pp. 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x
55. Linden van der W.J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, vol. 72, no 3, pp. 287–308. https://doi.org/10.1007/s11336-006-1478-z
56. Lo S., Andrews S. (2015) To Transform Or Not to Transform: Using Generalized Linear Mixed Models to Analyse Reaction Time Data. *Frontiers in Psychology*, vol. 6, August, Article no 1171. https://doi.org/10.3389/fpsyg.2015.01171
57. Lord F.M. (1953) The Relation of Test Score to the Trait Underlying the Test. *Educational and Psychological Measurement*, vol. 13, no 4, pp. 517–549. https://doi.org/10.1177/001316445301300401
58. Ludlow L.H. (1998) Galton: The First Psychometrician. *Popular Measurement*, vol. 1, no 1, pp. 13–14.
59. Maas van der H.L.J., Molenaar D., Maris G., Kievit R.A., Borsboom D. (2011) Cognitive Psychology Meets Psychometric Theory: On the Relation between Process Models for Decision Making and Latent Variable Models for Individual Differences. *Psychological Review,* vol. 118, no 2, pp. 339–356. https://doi.org/10.1037/a0022749
60. Meade A.W., Lautenschlager G.J. (2004) A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, vol. 7, no 4, pp. 361–388. https://doi.org/10.1177/1094428104268027
61. Miller G.A. (2003) The Cognitive Revolution: A Historical Perspective. *Trends in Cognitive Sciences*, vol. 7, no 3, pp. 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9
62. Molenaar D., Tuerlinckx F., van der Maas H.L. (2015) A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, vol. 50, no 1, pp. 56–74. https://doi.org/10.1080/00273171.2014.962684
63. Molenaar P.C.M., Beltz A.M. (2020) Modeling the Individual: Bridging Nomothetic and Idiographic Levels of Analysis. *The Cambridge Handbook of Research Methods in Clinical Psychology* (eds A.G.C. Wright, M.N. Hallquist), Cambridge: Cambridge University, pp. 327–336. https://doi.org/10.1017/9781316995808.031
64. Moore J. (1999) The Basic Principles of Behaviorism. *The Philosophical Legacy of Behaviorism* (ed. B.A. Thyer), Dordrecht: Springer Science+Business Media, pp. 41–68. https://doi.org/10.1007/978-94-015-9247-5_2
65. Moore J. (1996) On the Relation between Behaviorism and Cognitive Psychology. *The Journal of Mind and Behavior*, vol. 17, no 4, pp. 345–367.
66. Morís Fernández L., Vadillo M.A. (2020) Flexibility in Reaction Time Analysis: Many Roads to a False Positive? *Royal Society Open Science*, vol. 7, no 2, Article no 190831. https://doi.org/10.1098/rsos.190831
67. Parsons S., Kruijt A.-W., Fox E. (2019) Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measure-

ments. *Advances in Methods and Practices in Psychological Science*, vol. 2, no 4, pp. 378–395. https://doi.org/10.1177/2515245919879695
68. Passolunghi M.C., Siegel L.S. (2001) Short-Term Memory, Working Memory, and Inhibitory Control in Children with Difficulties in Arithmetic Problem Solving. *Journal of Experimental Child Psychology*, vol. 80, no 1, pp. 44–57. https://doi.org/10.1006/jecp.2000.2626
69. Pronk T., Hirst R.J., Wiers R.W., Murre J.M.J. (2023) Can We Measure Individual Differences in Cognitive Measures Reliably via Smartphones? A Comparison of the Flanker Effect across Device Types and Samples. *Behavior Research Methods*, vol. 55, no 4, pp. 1641–1652. https://doi.org/10.3758/s13428-022-01885-6
70. Putnick D.L., Bornstein M.H. (2016) Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review*, vol. 41, June, pp. 71–90. https://doi.org/10.1016/j.dr.2016.06.004
71. Rasch G. (1968) A Mathematical Theory of Objectivity and Its Consequences for Model Construction. *Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam*.
72. Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
73. Ratcliff R., Smith P.L., McKoon G. (2015) Modeling Regularities in Response Time and Accuracy Data with the Diffusion Model. *Current Directions in Psychological Science*, vol. 24, no 6, pp. 458–470. https://doi.org/10.1177/0963721415596228
74. Rey-Mermet A., Gade M., Oberauer K. (2018) Should We Stop Thinking about Inhibition? Searching for Individual and Age Differences in Inhibition Ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 44, no 4, pp. 501–526. https://doi.org/10.1037/xlm0000450
75. Rouder J.N., Haaf J.M. (2019) A Psychometrics of Individual Differences in Experimental Tasks. *Psychonomic Bulletin & Review*, vol. 26, no 2, pp. 452–467. https://doi.org/10.3758/s13423-018-1558-y
76. Rouder J., Kumar A., Haaf J.M. (2019) Why Most Studies of Individual Differences with Inhibition Tasks Are Bound to Fail. *PsyArXiv*. https://doi.org/10.31234/osf.io/3cjr5
77. Rousselet G.A., Wilcox R.R. (2020) Reaction Times and Other Skewed Distributions: Problems with the Mean and the Median. *Meta-Psychology*, vol. 4, Article no MP.2019.1630. https://doi.org/10.15626/MP.2019.1630
78. Royer J.M. (ed.) (2006) *The Cognitive Revolution on Educational Psychology*: *Current Perspectives on Cognition, Learning and Instruction.* Charlotte, NC: Information Age Publishing.
79. Saville C.W.N., Pawling R., Trullinger M., Daley D., Intriligator J., Klein C. (2011) On the Stability of Instability: Optimising the Reliability of Intra-Subject Variability of Reaction Times. *Personality and Individual Differences*, vol. 51, no 2, pp. 148–153. https://doi.org/10.1016/j.paid.2011.03.034
80. Scarpina F., Tagini S. (2017) The Stroop Color and Word Test. *Frontiers in Psychology*, vol. 8, April, Article no 557. https://doi.org/10.3389/fpsyg.2017.00557
81. Schmidt J.R., Besner D. (2008) The Stroop Effect: Why Proportion Congruent Has Nothing to Do with Congruency and Everything to Do with Contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 34, no 3, pp. 514–523. https://doi.org/10.1037/0278-7393.34.3.514
82. Schramm P., Rouder J.N. (2019) Are Reaction Time Transformations Really Beneficial? *PsyArXiv*. https://doi.org/10.31234/osf.io/9ksa6
83. Shichel I., Tzelgov J. (2018) Modulation of Conflicts in the Stroop Effect. *Acta Psychologica*, vol. 189, pp. 93–102. https://doi.org/10.1016/j.actpsy.2017.10.007

84. Simon H.A. (1979) Information Processing Models of Cognition. *Annual Review of Psychology*, vol. 30, no 1, pp. 363–396. https://doi.org/10.1146/annurev.ps.30.020179.002051
85. Smedt de B., Gilmore C.K. (2011) Defective Number Module or Impaired Access? Numerical Magnitude Processing in First Graders with Mathematical Difficulties. *Journal of Experimental Child Psychology*, vol. 108, no 2, pp. 278–292. https://doi.org/10.1016/j.jecp.2010.09.003
86. Sokal M.M. (1987) Psychological Testing and American Society 1890–1930. New Brunswick: Rutgers University.
87. Speelman C.P., McGann M. (2013) How Mean is the Mean? *Frontiers in Psychology*, vol. 4, July, Article no 451. https://doi.org/10.3389/fpsyg.2013.00451
88. Spence R., Owens M., Goodyer I. (2012) Item Response Theory and Validity of the NEO-FFI in Adolescents. *Personality and Individual Differences*, vol. 53, no 6, pp. 801–807. https://doi.org/10.1016/j.paid.2012.06.002
89. Stahl C., Voss A., Schmitz F., Nuszbaum M., Tüscher O., Lieb K., Klauer K.C. (2014) Behavioral Components of Impulsivity. *Journal of Experimental Psychology: General,* vol. 143, no 2, pp. 850–886. https://doi.org/10.1037/a0033981
90. Sternberg R.J. (1981) Testing and Cognitive Psychology. *American Psychologist*, vol. 36, no 10, pp. 1181–1189. https://doi.org/10.1037/0003-066X.36.10.1181
91. Stroop J.R. (1935) Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology,* vol. 18, no 6, pp. 643–662. https://doi.org/10.1037/h0054651
92. Tavakol M., Dennick R. (2011) Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, vol. 2, June, pp. 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd
93. Terman L.M. (1924) The Mental Test as a Psychological Method. *Psychological Review,* vol. 31, no 2, pp. 93–117. https://doi.org/10.1037/h0070938
94. Troyer A.K., Leach L., Strauss E. (2006) Aging and Response Inhibition: Normative Data for the Victoria Stroop Test. *Aging, Neuropsychology, and Cognition*, vol. 13, no 1, pp. 20–35. https://doi.org/10.1080/138255890968187
95. Tuerlinckx F., De Boeck P.D. (2005) Two Interpretations of the Discrimination Parameter. *Psychometrika*, vol. 70, no 4, pp. 629–650. https://doi.org/10.1007/s11336-000-0810-3
96. Watrin J.P., Darwich R. (2012) On Behaviorism in the Cognitive Revolution: Myth and Reactions. *Review of General Psychology*, vol. 16, no 3, pp. 269–282. https://doi.org/10.1037/a0026766
97. Watson J.B. (1913) Psychology as the Behaviorist Views It. *Psychological Review*, vol. 20, no 2, pp. 158–177. https://doi.org/10.1037/h0074428
98. Wells F.L. (1912) The Relation of Practice to Individual Differences. *The American Journal of Psychology*, vol. 23, no 1, pp. 75–88. https://doi.org/10.2307/1413115
99. Whelan R. (2008) Effective Analysis of Reaction Time Data. *The Psychological Record*, vol. 58, no 3, pp. 475–482. https://doi.org/10.1007/BF03395630
100. Wicherts J.M. (2016) The Importance of Measurement Invariance in Neurocognitive Ability Testing. *The Clinical Neuropsychologist*, vol. 30, no 7, pp. 1006–1016. https://doi.org/10.1080/13854046.2016.1205136
101. Wijsen L.D., Borsboom D., Alexandrova A. (2022) Values in Psychometrics. *Perspectives on Psychological Science*, vol. 17, no 3, pp. 788–804. https://doi.org/10.1177/17456916211014183
102. Willoughby M.T., Wirth R.J., Blair C.B. (2012) Executive Function in Early Childhood: Longitudinal Measurement Invariance and Developmental Change. *Psychological Assessment*, vol. 24, no 2, pp. 418–431. https://doi.org/10.1037/a0025779
103. Wissler C. (1901) The Correlation of Mental and Physical Tests. *The Psychological Review: Monograph Supplements,* vol. 3, no 6, pp. i–62. https://doi.org/10.1037/h0092995

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. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Baayen R.H., Milin P. (2010) Analyzing Reaction Times. *International Journal of Psychological Research*, vol. 3, no 2, pp. 12–28. https://doi.org/10.21500/20112084.807

Bindra D., Scheier I.H. (1954) The Relation between Psychometric and Experimental Research in Psychology. *American Psychologist,* vol. 9, no 2, pp. 69–71. https://doi.org/10.1037/h0062472

Birnbaum A. (1958) *On the Estimation of Mental Ability. Series Report no 15, Project no 7755–7723*. Texas: Randolph Air Force Base, TX USAF School of Aviation Medicine.

Bolsinova M., Tijmstra J. (2018) Improving Precision of Ability Estimation: Getting More from Response Times. *British Journal of Mathematical and Statistical Psychology*, vol. 71, no 1, pp. 13–38. https://doi.org/10.1111/bmsp.12104

Borsboom D. (2006) The Attack of the Psychometricians. *Psychometrika*, vol. 71, no 3, pp. 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom D., Kievit R.A., Cervone D., Hood S.B. (2009) The Two Disciplines of Scientific Psychology, or: The Disunity of Psychology as a Working Hypothesis. *Dynamic Process Methodology in the Social and Developmental Sciences* (eds J. Valsiner, P. Molenaar, M. Lyra, N. Chaudhary), New York, NY: Springer, pp. 67–97. https://doi.org/10.1007/978-0-387-95922-1_4

Braat M., Engelen J., van Gemert T., Verhaegh S. (2020) The Rise and Fall of Behaviorism: The Narrative and the Numbers. *History of Psychology*, vol. 23, no 3, pp. 252–280. https://doi.org/10.1037/hop0000146

Brauer M., Curtin J.J. (2018) Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items. *Psychological Methods*, vol. 23, no 3, pp. 389–411. https://doi.org/10.1037/met0000159

Brown J. (1992) *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890–1930*. Princeton, NJ: Princeton University.

Caruso J.C. (2004) A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *European Journal of Psychological Assessment*, vol. 20, no 3, pp. 166–171. https://doi.org/10.1027/1015-5759.20.3.166

Cattell J.M., Galton F. (1890) Mental Tests and Measurements. *Mind*, vol. os-XV, iss. 59, pp. 373–381. https://doi.org/10.1093/mind/os-XV.59.373

Corneille O., Mierop A., Unkelbach C. (2020) Repetition Increases Both the Perceived Truth and Fakeness of Information: An Ecological Account. *Cognition*, vol. 205, December, Article no 104470. https://doi.org/10.1016/j.cognition.2020.104470

Cronbach L.J. (1957) The Two Disciplines of Scientific Psychology. *American Psychologist,* vol. 12, no 11, pp. 671–684. https://doi.org/10.1037/h0043943

Cronbach L.J., Shavelson R.J. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, vol. 64, no 3, pp. 391–418. https://doi.org/10.1177/0013164404266386

Cunnings I. (2012) An Overview of Mixed-Effects Statistical Models for Second Language Researchers. *Second Language Research*, vol. 28, no 3, pp. 369–382. https://doi.org/10.1177/0267658312443651

De Boeck P., Jeon M. (2019) An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, vol. 10, February, Article no 102. https://doi.org/10.3389/fpsyg.2019.00102

Dietrich J.F., Huber S., Klein E., Willmes K., Pixner S., Moeller K. (2016) A Systematic Investigation of Accuracy and Response Time Based Measures Used to Index ANS Acuity. *PLoS ONE*, vol. 11, no 9, Article no e0163076. https://doi.org/10.1371/journal.pone.0163076

Dietrich J.F., Huber S., Nuerk H.-C. (2015) Methodological Aspects to Be Considered When Measuring the Approximate Number System (ANS): A Research Review. *Frontiers in Psychology*, vol. 6, March, Article no 295. https://doi.org/10.3389/fpsyg.2015.00295

Dodonova Yu.A., Dodonov Yu.S. (2013) Faster on Easy Items, More Accurate on Difficult Ones: Cognitive Ability and Performance on a Task of Varying Difficulty. *Intelligence*, vol. 41, no 1, pp. 1–10. https://doi.org/10.1016/j.intell.2012.10.003

Draheim C., Mashburn C.A., Martin J.D., Engle R.W. (2019) Reaction Time in Differential and Developmental Research: A Review and Commentary on the Problems and Alternatives. *Psychological Bulletin*, vol. 145, no 5, pp. 508–535. https://doi.org/10.1037/bul0000192

Draheim C., Tsukahara J.S., Martin J.D., Mashburn C.A., Engle R.W. (2021) A Toolbox Approach to Improving the Measurement of Attention Control. *Journal of Experimental Psychology: General,* vol. 150, no 2, pp. 242–275. https://doi.org/10.1037/xge0000783

Drew T., Vogel E.K. (2008) Neural Measures of Individual Differences in Selecting and Tracking Multiple Moving Objects. *The Journal of Neuroscience*, vol. 28, no 16, pp. 4183–4191. https://doi.org/10.1523/JNEUROSCI.0556-08.2008

Dunn T.J., Baguley T., Brunsden V. (2014) From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation. *British Journal of Psychology*, vol. 105, no 3, pp. 399–412. https://doi.org/10.1111/bjop.12046

Edwards J.R. (2001) Ten Difference Score Myths. *Organizational Research Methods*, vol. 4, no 3, pp. 265–287. https://doi.org/10.1177/109442810143005

Eide P., Kemp A., Silberstein R.B., Nathan P.J., Stough C. (2002) Test-Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change. *The Journal of Psychology*, vol. 136, no 5, pp. 514–520. https://doi.org/10.1080/00223980209605547

Embretson S. (1994) Applications of Cognitive Design Systems to Test Development. *Cognitive Assessment: A Multidisciplinary Perspective* (ed. C.R. Reynolds), New York, NY: Springer Science+ Business Media, pp. 107–135. https://doi.org/10.1007/978-1-4757-9730-5_6

Embretson S., Gorin J. (2001) Improving Construct Validity with Cognitive Psychology Principles. *Journal of Educational Measurement*, vol. 38, no 4, pp. 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x

Friedman N.P., Miyake A. (2017) Unity and Diversity of Executive Functions: Individual Differences as a Window on Cognitive Structure. *Cortex*, no 86, pp. 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Galton F. (1879) Psychometric Experiments. *Brain*, vol. 2, no 2, pp. 149–162. https://doi.org/10.1093/brain/2.2.149

Galton F. (1883) *Inquiries into Human Faculty and Its Development.* New York, NY: MacMillan. https://doi.org/10.1037/14178-000

Gevins A., Smith M.E. (2000) Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. *Cerebral Cortex*, vol. 10, no 9, pp. 829–839. https://doi.org/10.1093/cercor/10.9.829

Glaser R. (1981) The Future of Testing: A Research Agenda for Cognitive Psychology and Psychometrics. *American Psychologist*, vol. 36, no 9, pp. 923–936. https://doi.org/10.1037/0003-066X.36.9.923

Goldhammer F., Naumann J., Stelter A., Tóth K., Rölke H., Klieme E. (2014) The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights from a Computer-Based Large-Scale Assessment. *Journal*

of Educational Psychology, vol. 106, no 3, pp. 608–626. https://doi.org/10.1037/a0034716

Goldstein H. (2012) Francis Galton, Measurement, Psychometrics and Social Progress. *Assessment in Education: Principles, Policy & Practice*, vol. 19, no 2, pp. 147–158. https://doi.org/10.1080/0969594X.2011.614220

Goodhew S.C., Edwards M. (2019) Translating Experimental Paradigms into Individual-Differences Research: Contributions, Challenges, and Practical Recommendations. *Consciousness and Cognition*, vol. 69, January, pp. 14–25. https://doi.org/10.1016/j.concog.2019.01.008

Hambleton R.K. (1989) Principles and Selected Applications of Item Response Theory. *Educational Measurement* (ed. R.L. Linn), New York, NY: Macmillan Publishing Co, Inc; American Council on Education, pp. 147–200.

Heathcote A., Popiel S.J., Mewhort D.J. (1991) Analysis of Response Time Distributions: An Example Using the Stroop Task. *Psychological Bulletin,* vol. 109, no 2, pp. 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Heck D.W., Erdfelder E. (2016) Extending Multinomial Processing Tree Models to Measure the Relative Speed of Cognitive Processes. *Psychonomic Bulletin & Review*, vol. 23, no 5, pp. 1440–1465. https://doi.org/10.3758/s13423-016-1025-6

Hedge C., Powell G., Sumner P. (2018) The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behavior Research Methods*, vol. 50, no 3, pp. 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hepp H.H., Maier S., Hermle L., Spitzer M. (1996) The Stroop Effect in Schizophrenic Patients. *Schizophrenia Research*, vol. 22, no 3, pp. 187–195. https://doi.org/10.1016/S0920-9964(96)00080-1

Jensen A.R. (2006) *Clocking the Mind: Mental Chronometry and Individual Differences*. Amsterdam: Elsevier.

Johnson R.C., McClearn G.E., Yuen S., Nagoshi C.T., Ahern F.M., Cole R.E. (1985) Galton's Data a Century Later. *American Psychologist,* vol. 40, no 8, pp. 875–892. https://doi.org/10.1037/0003-066X.40.8.875

Kane M.J., Hambrick D.Z., Tuholski S.W., Wilhelm O., Payne T.W., Engle R.W. (2004) The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General,* vol. 133, no 2, pp. 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kievit R.A., Frankenhuis W.E., Waldorp L.J., Borsboom D. (2013) Simpson's Paradox in Psychological Science: A Practical Guide. *Frontiers in Psychology*, vol. 4, August, Article no 513. https://doi.org/10.3389/fpsyg.2013.00513

Kim E.S., Yoon M. (2011) Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 18, no 2, pp. 212–228. https://doi.org/10.1080/10705511.2011.557337

Kim S., Feldt L.S. (2010) The Estimation of the IRT Reliability Coefficient and Its Lower and Upper Bounds, with Comparisons to CTT Reliability Statistics. *Asia Pacific Education Review*, vol. 11, no 2, pp. 179–188. https://doi.org/10.1007/s12564-009-9062-8

Kleka P., Soroko E. (2018) How to Avoid the Sins of Questionnaire Abridgement — Guideline. *PsyArXiv*. https://doi.org/10.31234/osf.io/8jg9u

Lamiell J.T. (1992) Personality Psychology and the Second Cognitive Revolution. *American Behavioral Scientist*, vol. 36, no 1, pp. 88–101. https://doi.org/10.1177/0002764292036001008

Lazarsfeld P.F. (1950) The Logical and Mathematical Foundation of Latent Structure Analysis. *Studies in Social Psychology in World War II. Vol. IV: Measurement and Prediction* (eds S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld), Princeton: Princeton University, pp. 362–412.

Leitgöb H., Seddig D., Asparouhov T., Behr D., Davidov E., de Roover K. et al. (2023) Measurement Invariance in the Social Sciences: Historical Development, Methodological Challenges, State of the Art, and Future Perspectives. *Social Science Research*, vol. 110, January, Article no 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Linden van der W.J. (2009) Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, vol. 46, no 3, pp. 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Linden van der W.J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, vol. 72, no 3, pp. 287–308. https://doi.org/10.1007/s11336-006-1478-z

Lo S., Andrews S. (2015) To Transform Or Not to Transform: Using Generalized Linear Mixed Models to Analyse Reaction Time Data. *Frontiers in Psychology*, vol. 6, August, Article no 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lord F.M. (1953) The Relation of Test Score to the Trait Underlying the Test. *Educational and Psychological Measurement*, vol. 13, no 4, pp. 517–549. https://doi.org/10.1177/001316445301300401

Ludlow L.H. (1998) Galton: The First Psychometrician. *Popular Measurement*, vol. 1, no 1, pp. 13–14.

Maas van der H.L.J., Molenaar D., Maris G., Kievit R.A., Borsboom D. (2011) Cognitive Psychology Meets Psychometric Theory: On the Relation between Process Models for Decision Making and Latent Variable Models for Individual Differences. *Psychological Review,* vol. 118, no 2, pp. 339–356. https://doi.org/10.1037/a0022749

Meade A.W., Lautenschlager G.J. (2004) A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, vol. 7, no 4, pp. 361–388. https://doi.org/10.1177/1094428104268027

Miller G.A. (2003) The Cognitive Revolution: A Historical Perspective. *Trends in Cognitive Sciences*, vol. 7, no 3, pp. 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Molenaar D., Tuerlinckx F., van der Maas H.L. (2015) A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, vol. 50, no 1, pp. 56–74. https://doi.org/10.1080/00273171.2014.962684

Molenaar P.C.M., Beltz A.M. (2020) Modeling the Individual: Bridging Nomothetic and Idiographic Levels of Analysis. *The Cambridge Handbook of Research Methods in Clinical Psychology* (eds A.G.C. Wright, M.N. Hallquist), Cambridge: Cambridge University, pp. 327–336. https://doi.org/10.1017/9781316995808.031

Moore J. (1999) The Basic Principles of Behaviorism. *The Philosophical Legacy of Behaviorism* (ed. B.A. Thyer), Dordrecht: Springer Science+Business Media, pp. 41–68. https://doi.org/10.1007/978-94-015-9247-5_2

Moore J. (1996) On the Relation between Behaviorism and Cognitive Psychology. *The Journal of Mind and Behavior*, vol. 17, no 4, pp. 345–367.

Morís Fernández L., Vadillo M.A. (2020) Flexibility in Reaction Time Analysis: Many Roads to a False Positive? *Royal Society Open Science*, vol. 7, no 2, Article no 190831. https://doi.org/10.1098/rsos.190831

Parsons S., Kruijt A.-W., Fox E. (2019) Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, vol. 2, no 4, pp. 378–395. https://doi.org/10.1177/2515245919879695

Passolunghi M.C., Siegel L.S. (2001) Short-Term Memory, Working Memory, and Inhibitory Control in Children with Difficulties in Arithmetic Problem Solving. *Journal of Experimental Child Psychology*, vol. 80, no 1, pp. 44–57. https://doi.org/10.1006/jecp.2000.2626

Pronk T., Hirst R.J., Wiers R.W., Murre J.M.J. (2023) Can We Measure Individual Differences in Cognitive Measures Reliably via Smartphones? A Comparison of the Flanker Effect across Device Types and Samples. Behavior Research Methods, vol. 55, no 4, pp. 1641–1652. https://doi.org/10.3758/s13428-022-01885-6

Putnick D.L., Bornstein M.H. (2016) Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review*, vol. 41, June, pp. 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rasch G. (1968) A Mathematical Theory of Objectivity and Its Consequences for Model Construction. *Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam*.

Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Ratcliff R., Smith P.L., McKoon G. (2015) Modeling Regularities in Response Time and Accuracy Data with the Diffusion Model. *Current Directions in Psychological Science*, vol. 24, no 6, pp. 458–470. https://doi.org/10.1177/0963721415596228

Rey-Mermet A., Gade M., Oberauer K. (2018) Should We Stop Thinking about Inhibition? Searching for Individual and Age Differences in Inhibition Ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 44, no 4, pp. 501–526. https://doi.org/10.1037/xlm0000450

Rouder J.N., Haaf J.M. (2019) A Psychometrics of Individual Differences in Experimental Tasks. *Psychonomic Bulletin & Review*, vol. 26, no 2, pp. 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder J., Kumar A., Haaf J.M. (2019) Why Most Studies of Individual Differences with Inhibition Tasks Are Bound to Fail. *PsyArXiv*. https://doi.org/10.31234/osf.io/3cjr5

Rousselet G.A., Wilcox R.R. (2020) Reaction Times and Other Skewed Distributions: Problems with the Mean and the Median. *Meta-Psychology*, vol. 4, Article no MP.2019.1630. https://doi.org/10.15626/MP.2019.1630

Royer J.M. (ed.) (2006) *The Cognitive Revolution on Educational Psychology*: *Current Perspectives on Cognition, Learning and Instruction.* Charlotte, NC: Information Age Publishing.

Saville C.W.N., Pawling R., Trullinger M., Daley D., Intriligator J., Klein C. (2011) On the Stability of Instability: Optimising the Reliability of Intra-Subject Variability of Reaction Times. *Personality and Individual Differences*, vol. 51, no 2, pp. 148–153. https://doi.org/10.1016/j.paid.2011.03.034

Scarpina F., Tagini S. (2017) The Stroop Color and Word Test. *Frontiers in Psychology*, vol. 8, April, Article no 557. https://doi.org/10.3389/fpsyg.2017.00557

Schmidt J.R., Besner D. (2008) The Stroop Effect: Why Proportion Congruent Has Nothing to Do with Congruency and Everything to Do with Contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 34, no 3, pp. 514–523. https://doi.org/10.1037/0278-7393.34.3.514

Schramm P., Rouder J.N. (2019) Are Reaction Time Transformations Really Beneficial? *PsyArXiv*. https://doi.org/10.31234/osf.io/9ksa6

Schultz D.P., Schultz S.E. (1998) *Istoriya sovremennoy psikhologii* [A History of Modern Psychology]. Saint-Petersburg: Evraziya.

Shichel I., Tzelgov J. (2018) Modulation of Conflicts in the Stroop Effect. *Acta Psychologica*, 189, pp. 93–102. https://doi.org/10.1016/j.actpsy.2017.10.007

Simon H.A. (1979) Information Processing Models of Cognition. *Annual Review of Psychology*, vol. 30, no 1, pp. 363–396. https://doi.org/10.1146/annurev.ps.30.020179.002051

Smedt de B., Gilmore C.K. (2011) Defective Number Module or Impaired Access? Numerical Magnitude Processing in First Graders with Mathematical Difficulties. *Journal of Experimental Child Psychology*, vol. 108, no 2, pp. 278–292. https://doi.org/10.1016/j.jecp.2010.09.003

Sokal M.M. (1987) Psychological Testing and American Society 1890–1930. New Brunswick: Rutgers University.

Speelman C.P., McGann M. (2013) How Mean is the Mean? *Frontiers in Psychology*, vol. 4, July, Article no 451. https://doi.org/10.3389/fpsyg.2013.00451

Spence R., Owens M., Goodyer I. (2012) Item Response Theory and Validity of the NEO-FFI in Adolescents. *Personality and Individual Differences*, vol. 53, no 6, pp. 801–807. https://doi.org/10.1016/j.paid.2012.06.002

Stahl C., Voss A., Schmitz F., Nuszbaum M., Tüscher O., Lieb K., Klauer K.C. (2014) Behavioral Components of Impulsivity. *Journal of Experimental Psychology: General,* vol. 143, no 2, pp. 850–886. https://doi.org/10.1037/a0033981

Sternberg R.J. (1981) Testing and Cognitive Psychology. *American Psychologist*, vol. 36, no 10, pp. 1181–1189. https://doi.org/10.1037/0003-066X.36.10.1181

Stroop J.R. (1935) Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology,* vol. 18, no 6, pp. 643–662. https://doi.org/10.1037/h0054651

Tavakol M., Dennick R. (2011) Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, vol. 2, June, pp. 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Terman L.M. (1924) The Mental Test as a Psychological Method. *Psychological Review,* vol. 31, no 2, pp. 93–117. https://doi.org/10.1037/h0070938

Troyer A.K., Leach L., Strauss E. (2006) Aging and Response Inhibition: Normative Data for the Victoria Stroop Test. *Aging, Neuropsychology, and Cognition*, vol. 13, no 1, pp. 20–35. https://doi.org/10.1080/138255890968187

Tuerlinckx F., De Boeck P.D. (2005) Two Interpretations of the Discrimination Parameter. *Psychometrika*, vol. 70, no 4, pp. 629–650. https://doi.org/10.1007/s11336-000-0810-3

Voronin I.A., Zakharov I.M., Tabueva A.O., Merzon L.A. (2020) Diffuznaya model´ prinyatiya resheniya: otsenka skorosti i tochnosti otvetov v zadachakh vybora iz dvukh al´ternativ v issledovaniyakh kognitivnykh protsessov i sposobnostey [Diffuse Decision-Making Model: Assessment of the Speed and Accuracy of Answers in the Problems of Choosing from Two Alternatives in the Study of Cognitive Processes and Abilities]. *The Theoretical and Experimental Psychology,* vol. 13, no 2, pp. 6–23.

Watrin J.P., Darwich R. (2012) On Behaviorism in the Cognitive Revolution: Myth and Reactions. *Review of General Psychology*, vol. 16, no 3, pp. 269–282. https://doi.org/10.1037/a0026766

Watson J.B. (1913) Psychology as the Behaviorist Views It. *Psychological Review*, vol. 20, no 2, pp. 158–177. https://doi.org/10.1037/h0074428

Wells F.L. (1912) The Relation of Practice to Individual Differences. *The American Journal of Psychology*, vol. 23, no 1, pp. 75–88. https://doi.org/10.2307/1413115

Whelan R. (2008) Effective Analysis of Reaction Time Data. *The Psychological Record*, vol. 58, no 3, pp. 475–482. https://doi.org/10.1007/BF03395630

Wicherts J.M. (2016) The Importance of Measurement Invariance in Neurocognitive Ability Testing. *The Clinical Neuropsychologist*, vol. 30, no 7, pp. 1006–1016. https://doi.org/10.1080/13854046.2016.1205136

Wijsen L.D., Borsboom D., Alexandrova A. (2022) Values in Psychometrics. *Perspectives on Psychological Science*, vol. 17, no 3, pp. 788–804. https://doi.org/10.1177/17456916211014183

Willoughby M.T., Wirth R.J., Blair C.B. (2012) Executive Function in Early Childhood: Longitudinal Measurement Invariance and Developmental Change. *Psychological Assessment*, vol. 24, no 2, pp. 418–431. https://doi.org/10.1037/a0025779

Wissler C. (1901) The Correlation of Mental and Physical Tests. *The Psychological Review: Monograph Supplements,* vol. 3, no 6, pp. i–62. https://doi.org/10.1037/h0092995