Психология. Журнал Высшей школы экономики.
2013. Т. 10. № 4. С. 42–66.

КАЛИБРАЦИЯ ПРОГНОЗОВ В ЗАДАЧАХ БИНАРНОГО ВЫБОРА[1]

Д.П. КОЛЕСНИКОВА, А.В. БЕЛЯНИН

Колесникова Диана Павловна — стажер-исследователь лаборатории экспериментальной и поведенческой экономики Международного института экономики и финансов НИУ ВШЭ, бакалавр экономики. Сфера научных интересов: теория принятия решений, образование и образовательные технологии.
Контакты: dianakolesnikova@gmail.com

Белянин Алексей Владимирович — доцент и заместитель директора Международного института экономики и финансов НИУ ВШЭ, заведующий лабораторией экспериментальной и поведенческой экономики НИУ ВШЭ, старший научный сотрудник ИМЭМО РАН, PhD in Economics. Сфера научных интересов: экспериментальная экономика, теория принятия решений, теория игр, прикладная микроэкономика.
Контакты: icef-research@hse.ru

Резюме

В работе рассматривается проблема уверенности человека в принятых решениях. Исследования, посвященные калибрации прогнозов (соотношению частоты фактически правильных ответов с оценками уверенности человека в их правильности), продемонстрировали наличие эффекта «избыточной уверенности»: люди склонны завышать вероятность того, что их ответы или предсказания верны. Авторы предположили, что оценка уверенности человека в принятом решении и точность калибрации зависят от метода измерения уверенности. В эксперименте принимали участие школьники (N = 50) и студенты (N = 36). Они отвечали на вопросы на общую эрудицию, определяя степень своей уверенности в ответах по стандартной шкале, вербально оценивая уверенность в выбранном варианте, или же по дуплексной шкале, когда оценивается уверенность как в выбранном, так и в отвергнутом варианте. В заключение участники делали ставки на то, что выбранные ими альтернативы окажутся правильными. Результаты исследования

[1] Авторы выражают глубокую признательность редактору спецвыпуска О.А. Гулевич за детальное обсуждение и ценные комментарии, которые позволили значительно улучшить качество изложения. Все утверждения и заключения в статье остаются на совести авторов.

подтвердили поставленные гипотезы, в частности, наличие эффекта «избыточной уверенности» на российской выборке, причем она оказалась ниже в условиях дуплексных шкал. Мы объясняем это явление остаточной неопределенностью, которая не снимается при сравнении двух вариантов ответа в отличие от стандартной шкалы, где внимание респондента концентрируется на одном выбранном им варианте. Кроме того, зависимость между уровнями уверенности и качеством прогноза оказалась линейной убывающей в случае стандартной шкалы и параболической — в случае дуплексной шкалы. Таким образом, если респондент испытывает большие сомнения в выбранном варианте, то повышение уверенности в ответе по дуплексной шкале в среднем приводит к увеличению вероятности того, что ответ будет правильным, тогда как если респондент полагает, что он знает ответ, его уверенность чаще оказывается избыточной вне зависимости от способа ее измерения. Наконец, мы выяснили, что ставки служат более надежным предсказателем качества прогноза, чем вербальные оценки, для тех случаев, когда принятые решения были рациональными, а уровни уверенности — непротиворечивыми.

Ключевые слова: *бинарный выбор, калибрация прогноза, показатель Брайера, избыточная уверенность, дуплексная шкала, ставка.*

В процессе принятия любого решения важную роль играет уверенность в правильности выбранной альтернативы. Чем больше человек уверен в своем решении, тем большее влияние оно оказывает на его поведение и полученные результаты, поэтому чем точнее люди осознают меру собственного знания и незнания, тем более точными являются их решения и тем лучше ожидаемые результаты.

Вероятно, поэтому в последние десятилетия исследователи все чаще обращаются к проблеме уверенности человека в точности собственных оценок. Объектом оценок могут быть метеорологические прогнозы (Murphy, Winkler, 1984), решения в области медицины (Berner, Graber, 2008; Christensen-Szalanski, Bushyhead, 1981; Yates et al., 1998), психологии (Keren, 1991; Lichtenstein et al., 1977; Wright, Ayton, 1988; Yates, 1990) и в экономической сфере (Aukutsionek, Belianin, 2001; Bornstein, Zickafoose, 1999; Camerer, Lovallo, 1999; Foster, Vohra, 1999; Kalai et al., 1999; Kirchler, Maciejovsky, 2002; Koellinger et al., 2007). В самом деле, значительная часть экономических решений, включая потребительский выбор благ (например, туристической путевки или театрального представления), и практически все бизнес-решения (с кем заключать контракт, инвестировать ли в новые активы, проводить ли рекламную кампанию и др.) принимаются в условиях неопределенности, следовательно, чем точнее представления о ней, тем лучше должно было бы быть решение.

Адекватные суждения о собственных способностях предугадывать или оценивать эту неопределенность должны повышать качество принятых решений, в особенности в условиях реальных стимулов. Тем не менее это естественное предсказание сбывается далеко не всегда. Так, частота фактически правильных ответов и оценки уверенности человека в их правильности (Glaser, Weber, 2010; Wu et al., 2008) нередко

оказывается слабо связанными. В таких случаях говорят о плохой *калибрации прогноза*, которая чаще всего проявляется в «избыточной уверенности»: люди склонны завышать вероятность того, что их ответы или предсказания верны (Lichtenstein et al., 1977; Ronis, Yates, 1987).

Качество калибрации прогнозов, естественно, зависит от ряда индивидуальных и ситуативных факторов — сложности задания, наличия у человека специальных знаний, а также процедуры вынесения решения. Так, уровень «избыточной уверенности» возрастает со сложностью задания (Erev et al., 1994; Ferrell, McGoey, 1980; Gigerenzer, 1991; Lichtenstein, Fischhoff, 1977; Suantak et al., 1996). В то же время он снижается, когда решение принимают специалисты в данной области (Juslin, 1994; Murphy, Brown, 1985; Yates, Curley, 1985), а также при вынесении решения в группах (Allwood, Granhag, 1996).

Цель настоящей работы — изучение того, каким образом оценка уверенности и точность калибрации прогнозов зависят от способа измерения уверенности.

Чаще всего уверенность измеряется с помощью вербальных шкал. Шкала может быть усеченной, где респонденты отмечают свою уверенность в ответе в диапазоне значений от «случайный выбор» (уверенность 0.5) до «точно правильно» (уверенность 1); или же полной, где левой границей выступает суждение «точно неправильно» (уверенность 0 — Juslin et al., 2007). Использование полной шкалы приводит к снижению избыточной уверенности на большей части диапазона (единичного интервала), и лишь при самых высоких уровнях уверенности ее избыточность, измеренная по полной шкале, может оказаться выше (Juslin, Persson, 2002).

Усеченная шкала используется в тех случаях, когда респондент оценивает уверенность в правильности одной выбранной альтернативы (этот вариант оценки мы называем стандартной шкалой). Полную шкалу естественно применять в тех случаях, когда респондента просят оценить уверенность как в выбранном им, так и в отвергнутом им варианте ответа (оценка по дуплексной шкале), поскольку в этом случае респондент сравнивает оба варианта, один из которых точно неправилен. Заметим, что сам по себе выбор шкалы помещает респондента в разные контексты решения. В самом деле, стандартная шкала фокусирует внимание респондента на (реальном или мнимом) знании ответа, что неявно предполагает уверенность не ниже 0.5, и предполагает подтверждение принятого решения в условиях уже разрешившегося внутреннего конфликта (снятой неопределенности). Напротив, дуплексная шкала вновь возвращает испытуемого к задаче сравнения обеих альтернатив, даже несмотря на то, что решение уже принято, и тем самым «размазывает» или «распределяет» уверенность в выбранном ответе по полной шкале (от 0 до 1) вместо усеченной (от 0.5 до 1).

Эти соображения подсказывают, что уровень вербальной уверенности есть не некоторый абсолют, но мера, зависимая от шкалы. Мы полагаем, что *уровень уверенности зависит от вербальной шкалы, с помощью которой производится измерение: при*

измерении по дуплексной шкале (с оценкой уверенности как в принятом, так и в отвергнутом варианте), он окажется ниже, чем при измерении по стандартной шкале (с оценкой уверенности только в выбранном варианте ответа) (гипотеза 1).

Поскольку люди систематически переоценивают точность своих прогнозов, демонстрируя «избыточную уверенность», мы можем предположить, что *соответствие уверенности в точности ответа и фактической точности (калибрация прогноза) будет лучше при использовании дуплексной, чем при использовании стандартной шкалы* (гипотеза 2).

Вербальные шкалы оценок уверенности, однако, не являются единственной ее мерой: как минимум, не менее весомым «свидетельством» уверенности являются связанные с ней материально подкрепленные действия. В экономической литературе именно такие действия интерпретируются как *истинное* выражение предпочтений в отличие от словесного заявления или суждения, которые могут восприниматься как «несерьезные». К числу таких действий относятся материальные ставки на правильный ответ, которые могут привести как к выигрышу дополнительной суммы, так и к проигрышу имеющихся у него денег (или привлекательных для него призов).

Само по себе использование ставок в качестве меры уверенности не ново: так, Д. Цезарини с соавт. (Cesarini et al., 2006) сравнили уверенность с материальными ставками и без оных в лабораторных условиях при помощи метода доверительных интервалов, т.е. предварительно спросив респондентов, в каком интервале лежит их уровень уверенности в правильности данного ими ответа. Они обнаружили, что при использовании материальных стимулов избыточная уверенность снижается на 65%. П. Блаватский (Blavatskiy, 2008) применил несколько более сложную схему материального вознаграждения. Он предлагал участникам эксперимента с вопросами на общую эрудицию три варианта на выбор: (1) один из вопросов выбирается случайным образом, и испытуемый получает фиксированное вознаграждение (M), если ответ правильный; (2) проводится лотерея с фиксированным выигрышем (M) и вероятностью выигрыша, равной доле правильных ответов в общем количестве вопросов; (3) вариант (1) или (2) выбирается при помощи случайного устройства. Выбор первой опции означал, что испытуемый отличается избыточной уверенностью, выбор второй — недостаточной, третьей — его суждения хорошо откалиброваны. Все эти подходы, однако, не свободны от ограничений и недостатков: так, метод Д. Цезарини с соавт. содержит стратегические стимулы к завышению доверительных интервалов, а метод П. Блаватского не позволяет строить непрерывные меры калибрации.

В свете вышеизложенного мы полагаем, что *ставки на то, что данный ответ окажется правильным, окажутся тем более частыми и тем большими по величине, чем выше уровень вербально выраженной уверенности респондента в том, что он прав* (гипотеза 3). Кроме того, в соответствии с экономической логикой можно предположить, что *ставки*

будут лучше предсказывать фактическую точность ответов, чем вербально выраженная уверенность (гипотеза 4).

Для проверки этих гипотез было проведено эмпирическое исследование.

Метод исследования

Выборка. В исследовании приняли участие 86 респондентов, среди которых было 50 детей старшего школьного возраста (в среднем — 14 лет) и 36 студентов (в среднем — 19.6 года). 40% всех респондентов составляли юноши. Они представляли два образовательных учреждения города Рязань: среднюю школу и Рязанский государственный университет. Школьники заполняли методики на уроке информатики, где участие в эксперименте было организовано централизованно, но проходило добровольно: респонденты имели возможность отказаться. Студенты приглашались к участию в свободное время; набор осуществлялся при участии лидеров студенческих организаций.

Процедура исследования. Исследование проводилось в компьютерных классах с Интернет-соединением. Участники были рассажены за компьютерами и получали идентификационные номера. Экспериментатор объявлял о целях работы, добровольности участия, после чего озвучивал условия получения вознаграждения. Потом респонденты подписывали форму информированного согласия и могли приступить к выполнению задания.

Данные собирались при помощи онлайн-программного обеспечения (www.surveygizmo.com). Сначала респонденты в индивидуальном порядке отвечали на 45 вопросов на общую эрудицию, сразу после каждого вопроса на отдельном экране они оценивали свой уровень уверенности по вербальной шкале. Затем респондентам единым списком выводились все вопросы с вариантами ответов, и они получали возможность сделать ставки на правильность каждого ответа. Общее время работы составляло от 25 до 45 минут, причем в любой момент участники имели право задавать уточняющие вопросы. В заключение подсчитывались результаты и выдавалось вознаграждение.

Задание на общую эрудицию. Задание состояло из 45 вопросов по географии, истории и биологии, например: «Кто жил раньше? (а) Конфуций или (б) Аристотель» (см. приложение). На каждый из вопросов предлагались два варианта ответа, один из которых был объективно правильным. Участник должен был выбрать один из вариантов (тот, который, по его мнению, был правильным), а также оценить свою уверенность в данном ответе. Таким образом, в нашем эксперименте оценка уверенности по шкалам следовала за принятием решения (выбором одной из двух альтернатив) по каждому из вопросов и поэтому выступала не в качестве инструмента принятия решения, но как «портрет» ощущений респондента после того, как это решение было принято.

Вербальная шкала для оценки уверенности. Все участники были случайным образом разделены на две группы в зависимости от формата вербальной шкалы. Одна половина

участников получали анкету со стандартной шкалой от 0 — «случайный выбор» до 100 — «точно правильный ответ». С помощью этой шкалы они оценивали свою уверенность в точности выбранной альтернативы. Вторая половина респондентов получали анкету с дуплексной шкалой от 0 — «точно неправильный ответ» до 100 — «точно правильный ответ», по которой им следовало оценить выбранную и отвергнутую альтернативу. В этом случае отдельно оговаривалось, что уровни уверенности на обеих шкалах не обязательно в сумме дают 100.

Таким образом, шкалы различались по двум критериям. Основным критерием было количество альтернатив, которые оценивал участник (стандартная или дуплексная шкала). Дополнительным критерием были варианты ответов (полная или усеченная шкала). Дополнительный критерий различия был введен для того, чтобы избежать контринтуитивных значений уверенности. В самом деле, использование усеченной шкалы для дуплексной оценки привело бы к тому, что уровни уверенности по обеим шкалам не сходились бы к 1 ни в одном случае, кроме «углового» решения, когда оба уровня равны 0.5. Использование же полной шкалы для стандартной задачи оценки уверенности выбранного варианта чревато тем, что респонденты будут сообщать, что в выбранном варианте они уверены менее чем наполовину, что приводит к внутреннему противоречию.

Ставки на правильность ответа. Кроме оценочных мер собственной уверенности, в нашем эксперименте использовались материальные: после ответа на все вопросы респонденты могли сделать ставки на то, что их ответ на любой вопрос из анкеты окажется правильным. Иначе говоря, все решения приводили к ощутимыми материальным последствиям — получению или неполучению денежных или материальных призов.

Определение величины вознаграждения происходило следующим образом. Согласно начальным условиям, каждый участник получал 5000 условных единиц (у.е.) за сам факт участия. Этот выигрыш мог быть изменен за счет ставок. Минимальный размер ставки составлял 130 у.е., максимальный — 790 у.е. (границы определялись исходя из выигрыша в случае ожидаемого числа правильных ответов, а также ограничениями на желательное максимальное и минимальное количество ставок). Общая сумма ставок ограничивалась вознаграждением за участие, т.е. 5000 единиц. Экспериментатор много раз обращал внимание на это обстоятельство. В случае если ставка сделана на вопрос, на который испытуемый ответил правильно, он получал полный размер ставки, в противоположном случае терял половину своей ставки. Эта процедура, во-первых, позволяла ограничить число ставок, во-вторых, создавала достаточные стимулы для того, чтобы сделать ставки на те ответы, в правильности которых участник достаточно хорошо уверен.

С тем чтобы простимулировать участников давать наилучшие ответы, в эксперименте использовалось реальное вознаграждение. Баллы, полученные каждым участником по итогам ставок, в одних сессиях конвертировались в денежные выигрыши,

в других — в наборы фруктов. Школьники получали в качестве вознаграждений только наборы фруктов в соответствии с таблицей 1. Большинство студентов также получали только фрукты по той же шкале, однако каждый пятнадцатый участник студенческих групп отбирался случайным образом и получал возможность забрать свой выигрыш или в денежной форме по курсу 5 у.е. = 1 руб., или в виде фруктов (такая замена стимулов была продиктована бюджетными соображениями). С учетом заработанных баллов максимальный возможный денежный выигрыш в эксперименте составлял 2000 рублей — неплохой стимул для учащихся, так что все, кто получил такую возможность, предпочли деньги фруктам. Размер вознаграждения зависел от принятых во время эксперимента решений, а именно от размера ставок и от того, правильными ли оказались ответы в вопросах, на которые делались ставки.

Результаты исследования

Вербальная оценка уверенности и точность ответов

В таблице 2 представлены показатели уверенности в ответах и их фактической точности. Как видно из таблицы, средняя доля правильных ответов находится на уровне 55% и практически не зависит от используемой вербальной шкалы (что не удивительно, учитывая, что вопросы, касающиеся уверенности, задавались после ответов на вопросы анкеты). Напротив, степень уверенности зависит от вербальной шкалы: респонденты, оценивающие свою уверенность по стандартной шкале, оказываются значимо более уверенными, чем участники, оценивающие уверенность по дуплексной шкале ($t_{Стьюдента} = 5.61$, $p < 0.0000$; Wilcoxon—Mann—Whitney test $z = 4.68$, $p < 0.0000$).

Результат 1. При оценке уверенности методом дуплексных шкал уровни уверенности в среднем оказываются ниже, чем при ее оценке по стандартной шкале. Таким образом, использование дуплексной шкалы вместо стандартной снижает уровень уверенности в правильности выбранного варианта. Это позволяет утверждать, что наша гипотеза 1 в целом подтверждается.

Значимость этого факта в нынешней формулировке не следует преувеличивать, поскольку она связана со спецификой шкал: первая — усеченная, вторая — полная. На рисунке 1

Таблица 1

Соответствие условных единиц и призовых выигрышей

у.е.	ФРУКТЫ	
500	Киви	Мандарин
1000	Банан	Пластиковый стаканчик с виноградом
2000	Груша	Яблоко
5000	Грейпфрут	Свити
6000	Ананас	

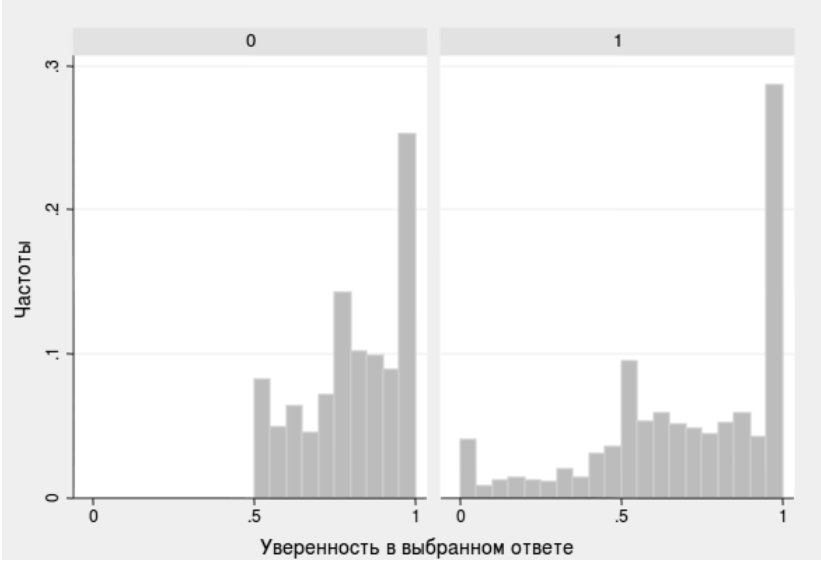
Таблица 2

Оценки уверенности и фактическая точность ответов

N	Всего		Шкала			
			Стандартная		*Дуплексная*	
	Точность	Уверенность	Точность	Уверенность	Точность	Уверенность
	86	86	43	43	43	43
Среднее	0.557	0.747	0.564	0.807	0.550	0.687
Медиана	0.556	0.759	0.556	0.827	0.556	0.674
Ст. откл.	0.075	0.115	0.083	0.074	0.067	0.118
Min	0.356	0.399	0.356	0.636	0.422	0.399
Max	0.733	0.975	0.733	0.975	0.689	0.902

Рисунок 1

Распределение уровней уверенности в правильности выбранного ответа (правый график (0) – контрольная, левый график (1) – экспериментальная группа)



представлены распределения уровней уверенности в правильности выбранных ответов для групп со стандартной и дуплексной шкалами. Как видно из этих графиков, усечение левого графика от 0.5 по сравнению с правым само по себе завышает средние значения в группе со стандартной шкалой, что не позволяет сравнивать их напрямую, поскольку часть ответов в экспериментальной группе лежит левее 0.5. Тем не менее

можно воспользоваться непараметрическими тестами, сравнив ранги уверенностей в ответах на каждый из 45 вопросов в контрольной (со стандартной шкалой) и экспериментальной (с дуплексной шкалой) подгруппах. Статистики Wilcoxon—Mann—Whitney для непарных выборок оказываются значимыми на 10%-ном уровне — для 25 вопросов (более половины), в том числе на 5%-ном уровне — для 21 вопроса; при этом во всех без исключения случаях большей была уверенность в вопросе со стандартной шкалой.

Насколько вербальная оценка уверенности соответствует точности ответов? Чтобы ответить на этот вопрос, был проведен анализ данных по нескольким направлениям.

Сначала мы сравнили *средние доли правильных ответов и уверенности респондентов в их правильности*. Доля правильных ответов в среднем оказывалась почти на 20 процентных пунктов ниже по сравнению с долей уверенности. Эта разница значима на любом уровне значимости для параметрических ($t_{Стьюдента} = -13.64, p < 0.0000$) и непараметрических (Wilcoxon—Mann—Whitney) двусторонних тестов на равенство средних[2] и подтверждается по отдельности для стандартной и дуплексной шкал. При этом, как видно из таблицы 2, точность ответов оказывается систематически (и значимо) ниже, но, как и следовало ожидать, не различается в зависимости от шкал оценки уверенности.

Результат 2. Сравнение средней частоты правильных ответов со средними значениями уверенности подтверждает эффект «избыточной уверенности», наблюдавшийся в предыдущих исследованиях.

При оценке по дуплексной шкале обращает на себя внимание наличие левого «хвоста», который «иррационален» в том смысле, что раз один из ответов точно правилен, степени уверенности должны суммироваться к единице. С формальной точки зрения, если респондент склоняется к одному из ответов, этот вариант должен получить больше 0.5; полное безразличие соответствует 0.5, а выбор варианта, в котором респондент уверен меньше чем наполовину, не имеет смысла.

В основном это правило выполнялось, однако случались и отклонения. При оценке по дуплексным шкалам в 10% случаев (202 раза) респонденты выбирали вариант, в котором они были уверены меньше, чем в отвергнутом. Ниже мы будем называть такие решения «иррациональными»; тут же отметим, что отклонения от рациональности встречаются нечасто и не носят систематического характера по индивидам: лишь у одного участника они встречаются в 30% случаев, у остальных же — гораздо реже.

Более существенно, что в 16% случаев респонденты выбирали вариант ответа, в котором были уверены менее чем на 50% (316 раз из 1906), причем максимальная доля таких ответов у одного респондента достигала 2/3, и ни у кого из респондентов не было так, чтобы они не

[2] В данном случае сравниваются средние из индивидуальных ответов, поэтому мы приводим статистику по Стьюденту, однако и WMW тест подтверждает ее результаты.

встречались ни разу. Проще всего было бы и эти последние случаи объявить иррациональными, однако не так все просто. Во-первых, с ними не связаны никакие систематические ошибки в ответах: доля правильных ответов для тех, кто выбирает вариант, в котором он уверен меньше, составляет 0.545, что статистически не отличается от общей средней ($t_{Стьюдента} = -0.88$, $p < 0.378$). Кроме того, из бесед с участниками экспериментов выяснилось, что некоторые из них избирали следующую стратегию: если человек совсем не знает, какой из вариантов правилен, для него естественнее отвечать не 0.5 в обоих случаях, а указать низкую уверенность и для одного, и для другого варианта. Усеченная шкала не дает возможности отследить этот уровень неуверенности и, соответственно, менее информативна в этом смысле, чем дуплексная.

То, что такое поведение имело место и приводило к снижению качества ответов, следует из сравнения результатов для тех случаев, когда респонденты называли такие уровни уверенности в ответах, что их сумма не сходилась к 1. Таких случаев набралось 772, или 40%, против 625 случаев (32%) когда сумма уверенностей превышала 1. И в том и в другом случае большинство ответов приходилось на небольшие отклонения, что позволяет трактовать их как случайные ошибки. Однако как раз среди тех ответов, сумма уверенностей в которых меньше 1, правильных оказалось значимо меньше: их доля составила 0.527 против 0.566 для остальных категорий (различие значимо на 10%-ном уровне, $t_{Стьюдента} = 1.67$, $p < 0.093$). Это наблюдение подсказывает, что низкие уровни уверенности в обоих вопросах чаще приводят к ошибкам, что является неожиданным и достаточно интересным результатом.

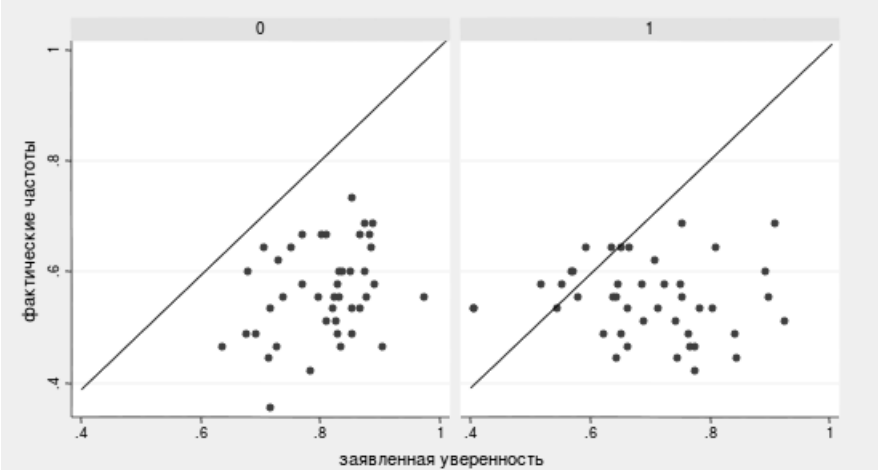
Результат 3. Низкие (менее 50%) уровни уверенности испытуемых в выбранных ответах, сообщенные в условиях дуплексных шкал, связаны с более частыми ошибками в ответах на вопросы.

Помимо сравнения средних долей правильных ответов и оценок уверенности респондентов в их правильности, интерес представляет более общий (интегральный) показатель, измеренный на всех интервалах уверенности, известный как калибрация прогнозов. Соотношения фактических частот (правильностей) и средних уверенностей представлены на рисунке 2, отдельно для респондентов со стандартной и дуплексной шкалами. Каждая точка соответствует средним величинам для одного индивида. На обоих графиках 45-градусная прямая соответствует ситуациям, когда фактическая частота правильных ответов совпадает с уверенностью. Большинство точек лежит правее и ниже этой прямой, что соответствует избыточной уверенности: ее заявленные уровни всегда выше фактических частот.

Это особенно отчетливо видно на примере группы со стандартной шкалой. Тем не менее именно для этой группы участников (и только для нее) наблюдается положительный тренд: чем выше заявленные уровни уверенности, тем больше и вероятность того, что ответ окажется правильным ($r = 0.15$, $p < 0.0000$). Вместе с тем корреляция между частотами правильных ответов и средней уверенностью, зафиксированной с

Рисунок 2

Калибрация ответов респондентов
(правый график (0) – контрольная, левый график (1) – экспериментальная группа)



помощью вербальной шкалы, оказывается слишком низкой, чтобы она могла считаться хорошим предсказателем угадывания.

Статистической мерой калибрации служит показатель Брайера (Brier, 1950), который вычисляется как

$$B = \frac{1}{N}\sum_{i=1}^{N}(r_i - c_i)^2, \qquad (1)$$

где N – общее число наблюдений, r_i – уверенность, а c_i – бинарная переменная, кодирующая фактическую правильность ответа. Чем выше этот показатель, тем сильнее среднее отклонение фактических частот от заявленных уровней уверенности, следовательно, тем хуже калибрация. В вопросах на общую эрудицию калибрация считается хорошей, если этот показатель не превышает 0.30 (Browne et al., 1999). В нашем случае при оценке на общем массиве данных он составляет 0.3057, т.е. лежит на верхней границе этой условной «нормы». При этом она оказывается несколько лучше для группы со стандартной шкалой (0.2868), чем для группы с дуплексной (0.3247) ($t_{Стьюдента} = -3.13$, $p < 0.0014$, WMW $z = -3.07$, $p < 0.0021$).

Результат 4. Показатель калибрации прогнозов оказывается лучшим для группы со стандартной, чем для группы с дуплексной шкалой. Таким образом, этот показатель свидетельствует против гипотезы 2.

Показатель Брайера сам по себе не очень содержателен – это всего лишь численная характеристика соответствий фактических и заявленных частот. Из содержательных соображений полезно посмотреть на его разложения, позволяющие статистически охарактеризовать причины отклонений правильности ответов от ожиданий респондента. Разложение

Мерфи (1973) представлено уравнением (2):

$$B = (1-\bar{c}) + \frac{1}{N}\sum_{t=1}^{T} n_t (r_t - \bar{c}_t)^2 - \\ -\frac{1}{N}\sum_{t=1}^{T} n_t (\bar{c}_t - \bar{c})^2. \qquad (2)$$

Здесь $\bar{c}$ – средняя доля правильных ответов для всей выборки, $\bar{c}_t$ – доля правильных ответов для каждой категории вероятностей (для наших целей мы используем 11 категорий: по одной на каждый десяток и 100 в качестве отдельной категории, поскольку таких ответов, как это обычно и бывает, в выборке достаточно много — 672, или 17%). Наконец, переменная n_t обозначает число наблюдений в каждой категории ответов, а r_t – заявленную вероятность по категориям.

Первое слагаемое показывает *вариацию исходов (outcome index variance, OIV)*, т.е. меру незнания респондентов, не связанную с их суждениями о собственном знании (или незнании). Третье слагаемое, называемое *разрешением Мерфи (Murphy resolution),* показывает степень различия правильности ответов от категории к категории, т.е. то, до какой степени аккуратность ответов различается в зависимости от уверенности респондентов в том, что они правы. Наконец, среднее слагаемое в уравнении, называемое *локальной надежностью (reliability-in-the-small, Rinsmall* — Yates, 1982), представляет собой усредненную по категориям меру различий между уверенностью и фактическими частотами, т.е. собственно калибрацию. Это слагаемое также представляет собой взвешенную сумму квадратов, поэтому, чтобы найти средневзвешенную меру несоответствия (число от 0 до 1), следует использовать корень квадратный из этой величины.

Аналогична размерность еще одного показателя — *глобальной надежности (reliability-in-the-large, Rinlarge)*, который можно вычленить из другого разложения показателя Брайера (Yates, 1982):

$$B = Var(c) + Var(r) + (\bar{r} - \bar{c})^2 - \\ -2 \cdot Cov(c,r), \qquad (3)$$

где $Var(c) = c(1 - c)$, а локальная надежность разложима как:

$$\frac{1}{N}\sum_{t=1}^{T} n_t (r_t - \bar{c}_t)^2 = Var(r) + (\bar{r} - \bar{c})^2 - \\ -2 \cdot Cov(r,c) + \frac{1}{N}\sum_{t=1}^{T} n_t (\bar{c}_t - \bar{c})^2. \qquad (4)$$

Второе слагаемое в этом выражении и есть глобальная надежность, которая в отличие от локальной показывает не средние расхождения уверенности и фактических частот, а то, до какой степени *средняя* уверенность отличается от средних же частот: чем выше эта величина, тем сильнее означенное различие, тем хуже калибрация и тем, соответственно, меньше оснований доверять прогнозу.

В нашем случае почти 80% показателя Брайера приходится на первое слагаемое (0.2466), что свидетельствует о том, что основную роль в его вариации играет незнание респондентов правильных ответов, а не локальная надежность (калибрация) этих оценок (она составляет лишь 0.067, остальное приходится на разрешение Мерфи). Таким образом,

значения показателя Брайера определяются, прежде всего, высокой вариацией незнания респондентов и в меньшей степени – переоценкой собственных знаний.

Эти оценки, однако, строго говоря, не обоснованы статистически, ибо не учитывают зависимости оценок, данных одним и тем же индивидом. В таблице 3 представлены статистики показателя Брайера и его разложений, подсчитанные отдельно для каждого респондента. Как следует из таблицы, и здесь показатели дисперсии исходов OIV составляют от 75 до 80% показателя Брайера, что подтверждает ранее сделанный вывод о том, что основной вклад в низкие показатели калибрации вносит незнание респондентов, а не переоценка прогностических способностей. Этот вывод, однако, не стоит абсолютизировать, поскольку ни показатель Брайера, ни его разложения не выявляют причин ошибок прогнозов – они лишь показывают, какой вклад вносят в эти ошибки разные статистические характеристики суждений.

Для более содержательного анализа сопоставим собственно калибрацию — локальную надежность и фактические частоты правильных ответов по экспериментальным условиям.

Таблица 3

Статистики индивидуальных показателей Брайера

Показатель	Brier	OIV	Murphy	Rinsmall	Rinlarge
Все (N = 86)					
Среднее	0.306	0.241	0.025	0.289	0.172
Медиана	0.305	0.247	0.020	0.290	0.168
Ст. откл.	0.057	0.011	0.017	0.081	0.099
Min	0.189	0.196	0	0.089	0.002
Max	0.495	0.250	0.074	0.472	0.391
Контрольная группа (N = 43)					
Среднее	0.287	0.239	0.022	0.252	0.201
Медиана	0.274	0.244	0.016	0.243	0.2
Ст. откл.	0.051	0.013	0.017	0.077	0.089
Min	0.189	0.196	0	0.089	0.024
Max	0.427	0.250	0.074	0.423	0.391
Экспериментальная группа (N = 43)					
Среднее	0.325	0.243	0.027	0.325	0.142
Медиана	0.309	0.247	0.031	0.305	0.12
Ст. откл.	0.058	0.009	0.016	0.068	0.1
Min	0.236	0.214	0	0.220	0.002
Max	0.495	0.250	0.062	0.472	0.389

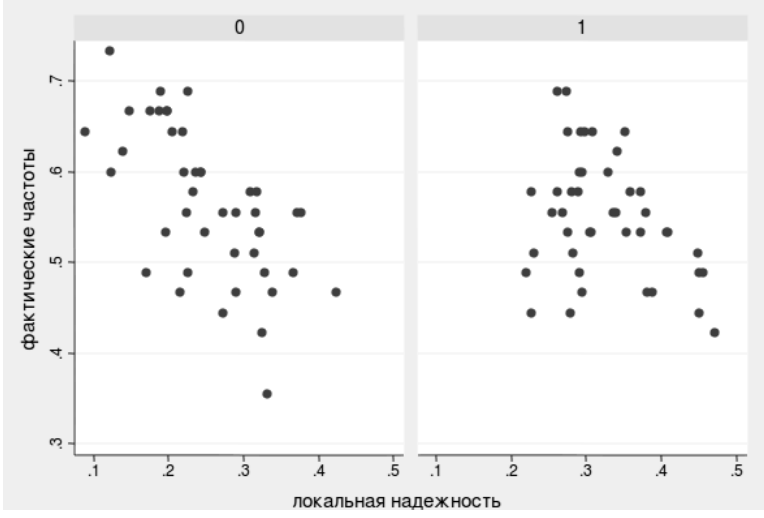
На рисунке 3 по вертикальной оси представлены доли правильных ответов, по горизонтальной — корень квадратный из локальной надежности оценок (Rinsmall); левый график соответствует группе со стандартной, а правый — группе с дуплексной шкалой.

Как следует из рисунка, для группы со стандартной шкалой наблюдается хоть и не сильная, но отчетливая отрицательная связь: чем выше показатель локальной надежности (т.е. чем хуже собственно калибрация), тем менее точными получаются ответы. Этот факт может служить количественным выражением расхождения между уровнями уверенности и фактическими частотами, которое возрастает по мере роста уверенности. Напротив, для группы с дуплексной шкалой зависимость скорее параболическая: для индивидов с низким уровнем локальной надежности (примерно до уровня в 0.35) рост этого показателя означает *рост* правильности ответов. Иными словами, даже несмотря на то, что калибрация прогнозов ухудшается с ростом уверенности, *при достаточно низких ее значениях возрастает локальная надежность*, т.е. большая уверенность означает большую точность суждения.

Иными словами, большая уверенность в своих суждениях в группе с дуплексной шкалой положительно связана с качеством ответов. Однако при такой шкале эта положительная зависимость, по-видимому, сохраняется только до тех пор, пока уверенность компенсируется переходом от полного незнания респондента к некоторому представлению о предмете вопроса. Это несколько неожиданное наблюдение свидетельствует о том, что соотношения между шкалами измерений и уровнями уверенности

Рисунок 3

Показатели правильности ответов и локальной надежности (правый график (0) — контрольная, левый график (1) — экспериментальная группа)



устроены более сложно, чем предполагалось в нашей гипотезе 2, и даже содержат элементы перформативности. Сформулируем их в качестве отдельного результата.

Результат 5. В стандартных задачах с одной шкалой точность ответов равномерно снижается по мере уменьшения калибрации. В задачах с дуплексной шкалой эта зависимость параболическая: при низкой калибрации ее рост связан с повышением качества прогнозов, а при высокой — с его снижением.

Денежные ставки и точность ответов

Ставка на то, что ответ правилен, представляет наибольший интерес с точки зрения экономики: ведь именно в этом случае респондент подтверждает свою уверенность материально мотивированным поступком. Корреляции величин и количеств ставок, сделанных одним участником, с их уровнями уверенности представлены в таблице 4.

Корреляции между уверенностью и величинами ставок оказались положительными и значимыми, что свидетельствует в пользу нашей гипотезы 3. Корреляции числа ставок не столь однозначны. Участники экспериментов со стандартными шкалами в среднем делали тем больше ставок, чем более были уверены в их правильности, хотя зависимость эта оказалась не очень сильной и значимой лишь на 10%-ном уровне. Напротив, участники экспериментов с дуплексными шкалами в среднем ставили тем *реже*, чем выше были их заявленные уровни уверенности, при том что размер этих ставок оказывался тем *больше*, чем выше была уверенность. Этот результат естественно интерпретировать следующим образом. В условиях дуплексной шкалы участники лишний раз взвешивают аргументы за и против каждого из вариантов ответа и материально подтверждают свой выбор только тогда, когда они действительно уверены в ответе, т.е. в *меньшем* количестве случаев, при том что размер ставки при этом оказывается *большим*. Иными словами, ставки при этом делаются реже, но только в тех случаях, когда уверенность в ответе достаточно высока. Тем самым наша гипотеза 3 подтверждается в следующем виде:

Результат 6. Уровень уверенности положительно связан с размерами ставок для всех экспериментальных условий, однако в условиях дуплексной шкалы приводит к тому, что ставки чаще делаются в тех случаях, когда респонденты сильнее всего уверены в собственных ответах.

Таблица 4

Корреляции уровней уверенности и показателей ставок

	Величина ставок	**Количество ставок**
Стандартная шкала	0.3830***	0.0408*
Дуплексная шкала	0.2470***	−0.084***

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Обратимся теперь к нашей гипотезе 4 и сравним влияние ставок и уровней уверенности на правильность ответов при дуплексной шкале. Как можно ожидать в свете предыдущего анализа, ставки влияют сильнее, чем вербальные оценки. И действительно, оценки пробит-модели зависимости правильности от уверенности и ставок показывают, что увеличение размера ставки на 50 у.е. в среднем повышает вероятность того, что ответ будет правильным, примерно на 1%, тогда как эффект вербальной уверенности не значим.

Наши данные, однако, позволяют пойти дальше и посмотреть, как связаны ставки и вербальная уверенность с точностью прогнозов в зависимости от того, сходятся или не сходятся уровни уверенности к 1, а также для рациональных и иррациональных решений (эти последние, напомним, соответствуют тем 10% ситуаций, когда респондент выбирает вариант ответа, в котором он менее уверен, по вербальной шкале). В таблице 5 представлены коэффициенты пробит-регрессий для правильности ответов в зависимости от соотношений вероятностей и рациональности ставок (контроль на уровень предпочтений в отношении риска не меняет качественной картины, представленной в таблице). Ответы разбиты на 6 категорий в соответствии с тем, была ли выбрана та альтернатива, в которой респондент был уверен больше (в этом случае будем говорить, что респондент делает рациональный выбор), или же другая (в этом случае его выбор нерационален), и параллельно в соответствии с тем, была ли сумма

Таблица 5

Средние предельные эффекты уверенности и ставки для правильности ответов

	Рациональный выбор					
	P(A) + P(B) = 1		P(A) + P(B) < 1		P(A) + P(B) > 1	
N	461		684		546	
	У	С	У	С	У	С
Предельный эффект	0.167	0.017***	−0.209	0.016	−0.074	0.006
Стандартное отклонение	0.109	0.004	0.168	0.016	0.235	0.005
	Нерациональный выбор					
	P(A) + P(B) = 1		P(A) + P(B) < 1		P(A) + P(B) > 1	
N	35		88		79	
	У	С	У	С	У	С
Предельный эффект	0.740*	0.021	0.776*	0.021	−0.867*	0.005
Стандартное отклонение	0.425	0.015	0.417	0.028	0.508	0.015

Примечание. По столбцам — категории рациональности выбора (бóльшая или меньшая уверенность в выбранном ответе), и сумма уверенностей, равная, меньшая или большая 1. По строкам — объясняющие переменные. У — уверенность, С — ставка. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

уверенностей равна 1, меньше 1 или больше 1. В качестве исследуемых объясняющих переменных использованы уровни уверенности в выбранном варианте ответа (нормированные к 1) и ставки, измеренные в десятках рублей (Wilcoxon—Mann—Whitney) (50 у.е.).

Как следует из таблицы, предельные эффекты невелики во всех случаях, однако характерно, что уверенность значимым образом связана с вероятностью правильного ответа ровно в тех случаях, когда выбор респондента не рационален (последние три столбца таблицы). Более того, средние предельные эффекты оказались положительными для тех случаев, когда суммы уверенностей не превышают единицы, тогда же, когда эта сумма превышена (в последнем столбце), этот эффект становится отрицательным. Эти выводы согласуются с рисунком 3 и результатом 5: в условиях дуплексной шкалы при низких уровнях уверенности ее рост означает повышение качества прогноза, тогда как при высоких уровнях уверенности — снижение этого качества.

Что касается ставок, то они оказываются значимыми в одном случае: когда выбор делается рациональным образом, а суммы уверенностей сходятся к 1. При этих условиях эффект ставки оказывается даже сильнее, чем для всей совокупности: ее увеличение на 50 у.е. в среднем приводит к повышению вероятности правильного ответа на 1.6%. Таким образом, сам факт ставки как мотивированного действия является более надежным «предсказателем» правильности ответа, чем вербальная уверенность, однако только в том случае, когда респондент ведет себя рационально и судит непротиворечиво. Наша гипотеза 4 получает подтверждение в более развернутом варианте:

Результат 7. Величины ставок являются значимым индикатором правильности ответа в тех и только в тех случаях, когда ответы рациональны, а уровни уверенности не противоречивы. Уровни вербальной уверенности являются значимыми предсказателями правильности для нерациональных ответов, причем влияют положительно на правильность при низких уровнях уверенности и отрицательно — при высоких.

Обсуждение и выводы

Мы рассмотрели влияние способов измерения уверенности на оценку уверенности и точность калибрации прогнозов. Результаты частично подтвердили поставленные гипотезы и позволили сделать ряд дополнительных заключений.

Вывод о том, что калибрация прогнозов далека от идеала, разумеется, не нов. Однако мы получили значимое подтверждение того, что сама ее величина зависит от способа измерения, причем зависит неоднозначным образом. С одной стороны, дуплексные шкалы приводят к значимо более низким показателям уверенности, так что их использование вроде бы должно снижать избыточную уверенность и улучшать калибрацию (Mussweiler, Posten, 2012). Однако этого не происходит: результат 4 прямо свидетельствует о том, что калибрация оказалась выше в случае стандартной шкалы.

Этому факту можно предложить два объяснения. С одной стороны,

для дуплексных шкал характерна высокая доля (40%) ответов, когда сумма уровней уверенности в двух вопросах меньше 1, что служит обоснованным индикатором незнания респондента (результат 3). С другой стороны (результат 5), локальная надежность, или показатель собственно калибрации, линейным образом снижается с долей правильных ответов при стандартной шкале, но демонстрирует параболическую взаимозависимость с долей правильных ответов. Это означает, что дуплексный метод может быть применен как мера борьбы с избыточной уверенностью в тех случаях, когда респондент обладает некоторыми знаниями, но скорее не уверен в том, что его знания точны. Эта гипотеза заслуживает самостоятельного исследования.

Результат 5 для группы с дуплексной шкалой представляет самостоятельный интерес еще и потому, что он может означать различие механизмов самооценки респондента в условиях, когда с ответом на вопрос в сознании ассоциируется разное количество степеней свободы. При всем формальном сходстве принципиальная разница между стандартной и дуплексной шкалами состоит в том, что при стандартной формулировке человека просят оценить степень своей уверенности в решении, которое он уже принял, вызывая тем самым целый спектр феноменов и коннотаций, не имеющих отношения к самому процессу выбора. Это и базовый уровень уверенности, связанный с темпераментом и/или уровнем притязаний, и психологическая потребность обосновать и оправдать собственное решение, и интуитивное ощущение меры собственного незнания (как правило, заниженное). Напротив, дуплексная шкала не предполагает никакого предпочтения в отношении двух вариантов: таким образом, еще не снимается психологическая неопределенность относительно ответа. Если респондент еще не снял для себя эту неопределенность, его сознание скорее будет занято не оправданием собственного выбора, а поиском аргументов в пользу правильного ответа. В этих условиях, если респондент на самом деле не принял для себя решение о том, какой из двух вариантов ответа правилен, то дуплексная шкала «регистрирует» это его состояние неопределенности, в котором он тщательнее взвешивает все за и против каждого из вариантов и в результате выбирает тот вариант, в пользу которого оказалось больше аргументов. Если же уверенность достаточно высока, то сознание «переключается» с задачи поиска ответа на задачу подтверждения принятого решения, где и сказывается завышенность собственных прогностических способностей. Для более тщательной проверки этого объяснения требуются дополнительные исследования, однако сама постановка вопроса о том, всегда ли человеку следует стремиться к скорейшему устранению избыточности контекста, представляет интерес с точки зрения теории и практики принятия решений.

Параболическая зависимость получает подтверждение и с другой стороны: в случаях, когда респонденты выбирают не тот вариант ответа, в котором они сами уверены больше, уровни уверенности служат значимым

предсказателем правильности ответа (результат 7, вторая часть). С содержательной точки зрения этот вывод означает, что если человек не знает, какой вариант ответа правилен, то его вербальная уверенность может быть даже более надежным индикатором правильности суждения, чем фактически сделанный выбор или ставки.

С другой стороны (результат 7, первая часть), мы получили сильную положительную зависимость между качеством ответов и ставками (мотивированным суждением) в тех и только тех случаях, когда выбор рационален, а уровни уверенности непротиворечивы на дуплексных шкалах. В этих случаях ставки служат более надежным индикатором, чем вербальная уверенность, и такие решения соответствуют логике рационального выбора с экономической точки зрения.

Для подтверждения надежности этих результатов необходимы дальнейшие исследования, однако если они подтвердятся, то можно будет делать вывод о том, что механизмы (а следовательно, и качество) принятия решений различаются у человека в условиях низкой и высокой уверенности в правильности ответа. Низкая уверенность означает неопределенность суждения, и поэтому принятые решения нередко оказываются иррациональными и «случайными». В этом случае вербальная оценка как более непосредственная и «интуитивная» оказывается точнее, чем само решение, даже если оно подкреплено материально. Напротив, у респондентов, которые принимают рациональные решения, работает стандартная экономическая логика, и для них действия (ставки) говорят больше, чем слова. Можно было бы утверждать, что это разбивает людей на три класса: «рационалистов», действия которых непротиворечивы и рациональны; «интуитивистов», действия которых противоречат их словам, однако слова говорят вернее действий; и тех, кто принимает рациональные решения, однако чьи вербальные уровни уверенности не согласуются со свойствами вероятностей. Построение подобной классификации позволило бы пролить свет на механизмы принятия решений и факторы, приводящие к повышению их качества.

Литература

Allwood C.M., Granhag P.A. Realism in confidence judgments as a function of working in dyads or alone // Organizational Behavior and Human Decision Processes. 1996. 66. 3. 277–289.

Aukutsionek S.P., Belianin A.V. Quality of forecasts and business performance: A survey study of Russian managers // Journal of Economic Psychology. 2001. 22. 5. 661–692.

Berner E.S., Graber M.L. Overconfidence as a cause of diagnostic error in medicine // The American Journal of Medicine. 2008. 121 (5 Suppl). S2–S23.

Blavatskiy P. Betting on own knowledge: Experimental test of overconfidence // University of Zurich WP 358, 2008.

Bornstein B.H., Zickafoose D.J. I know I know it, I know I saw it: The stability of overconfidence across domains // Journal of Experimental Psychology: Applied. 1999. 5. 1–13.

Camerer C., Lovallo D. Overconfidence and excess entry: An experimental approach // American Economic Review. 1999. 89. 1. 306–318.

Cesarini D., Sandewall Ö., Johannesson M. Confidence interval estimation tasks and the economics of overconfidence // Journal of Economic Behavior and Organization. 2006. 61. 3. 453–470.

Christensen-Szalanski J.J., Bushyhead J.B. Physicians' use of probabilistic information in a real clinical setting // Journal of Experimental Psychology: Human Perception and Performance. 1981. 7. 928–935.

Erev I., Wallsten T.S., Budescu D.V. Simultaneous over-and undercon dence: The role of error in judgment processes // Psychological Review. 1994. 101. 3. 519–527.

Ferrell W.R., McGoey P.J. A model of calibration for subjective probabilities // Organizational Behavior and Human Performance. 1980. 26. 1. 32–53.

Foster D.P., Vohra R. Regret in the on-line decision problem // Games and Economic Behavior. 1999. 29. 1. 7–35.

Gigerenzer G. How to make cognitive illusions disappear: Beyond «heuristics and biases» // European Review of Social Psychology. 1991. 2. 1. 83–115.

Glaser M., Weber M. Overconfidence // H.K. Baker, J.R. Nofsinger (eds.) Behavioral Finance: Investors, Corporations, and Markets. N.Y.: Wiley, 2010. P. 241–258.

Juslin P. The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items // Organizational Behavior and Human Decision Processes. 1994. 57. 2. 226–246.

Juslin P., Persson M. PROBabilities from EXemplars (PROBEX): A «lazy» algorithm for probabilistic inference from generic knowledge // Cognitive Science. 2002. 26. 5. 563–607.

Juslin P., Winman A., Hansson P. The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals // Psychological Review. 2007. 114. 3. 678–703.

Kalai E., Lehrer E., Smorodinsky R. Calibrated forecasting and merging // Games and Economic Behaviour. 1999. 29. 151–169.

Keren G. Calibration and probability judgements: Conceptual and methodological issues // Acta Psychologica. 1991. 77. 3. 217–273.

Kirchler E., Maciejovsky B. Simultaneous over-and underconfidence: Evidence from experimental asset markets // Journal of Risk and Uncertainty. 2002. 25. 1. 65–85.

Koellinger P., Minniti M., Schade C. «I think I can, I think I can»: Overconfidence and entrepreneurial behavior // Journal of Economic Psychology. 2007. 28. 4. 502–527.

Lichtenstein S., Fischhoff B. Do those who know more also know more about how much they know? // Organizational Behavior and Human Performance. 1977. 20. 2. 159–183.

Lichtenstein S., Fischhoff B., Phillips L.D. Calibration of probabilities: The state of the art. The Netherlands: Springer, 1977. P. 275–324.

Murphy A.H., Brown B.G. A comparative evaluation of objective and subjective weather forecasts in the United States // Behavioral decision making. N.Y.: Springer, 1985. P. 329–359.

Murphy A.H., Winkler R.L. Probability forecasting in meteorology // Journal of the American Statistical Association. 1984. 79. 387. 489–500.

Mussweiler T., Posten A.C. Relatively certain. Comparative thinking reduces uncertainty // Cognition. 2012. 122. 2. 236–240.

Ronis D.L., Yates J.F. Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method // Organizational Behavior and Human Decision Processes. 1987. 40. 193–218.

Suantak L., Bolger F., Ferrell W. The «hard-easy effect» in subjective probability calibration // Organizational Behavior and Human Decision Processes. 1996. 67. 201–221.

Wright G., Ayton P. Decision time, subjective probability, and task difficulty // Memory and Cognition. 1988. 16. 2. 176–185.

Wu S.W., Johnson J.E., Sung M.C. Overconfidence in judgements: the evidence, the implications and the limitations // The Journal of Prediction Markets. 2008. 2. 1. 73–90.

Yates J.F. Judgment and decision making. NJ: Prentice-Hall, 1990.

Yates J.F., Curley S.P. Conditional distribution analyses of probabilistic forecasts // Journal of Forecasting. 1985. 4. 1. 61–73.

Yates J.F., Lee J.-W., Shinotsuka H., Patalano A.L., Sieck W.R. Cross-cultural variations in probability judgment accuracy: beyond general knowledge overconfidence? // Organizational Behaviour and Human Decision Processes. 1998. 74. 89–117.

Приложение

Анкета для измерения общей эрудиции

Вам предстоит ответить на 45 тестовых вопросов на общие знания. Если вы не знаете, какой вариант ответа правильный, выберите тот, в котором вы уверены в большей степени. Время ответа на каждый вопрос не ограничено. Справа обозначьте, насколько вы уверены в своем ответе. Если вы выбираете крайние случаи, когда для вас это «случайный выбор» или когда вы знаете, что это «точно правильный ответ», обведите концы отрезка в кружочек. Когда же степень вашей уверенности находится между концами отрезка, поставьте засечку на линии.

1. Кто жил раньше: (А) Конфуций или (Б) Аристотель?
2. При ком появились первые бумажные деньги в России: (А) при Екатерине II или (Б) при Петре I?
3. У чего выше калорийность: (А) у стакана молока (3.2%) или (Б) у банана?
4. Перикл был правителем: (А) Спарты или (Б) Афин?
5. Авария на Чернобыльской АЭС произошла: (А) в 1986 или (Б) в 1989?
6. Кто из двух известных художников стал заниматься живописью только в зрелом возрасте (в 30 лет): (А) Ван Гог или (Б) Пабло Пикассо?
7. Где выше средняя температура января: (А) в Лондоне или (Б) в Киеве?
8. Чье знаменитое плавание завершилось раньше: (А) Васко да Гама или (Б) Магеллана?
9. Чья максимальная скорость выше: (А) первого автомобиля или (Б) бегемота?
10. Какая страна больше по площади: (А) Сан-Марино или (Б) Аргентина?
11. Что больше: (А) sin(1) или (Б) cos(1)?
12. Ревель – это старое русское название: Риги (А) или Таллина (Б)?
13. Какая страна больше по площади: (А) Португалия или (Б) Азербайджан?
14. Какая страна больше по площади: (А) Сан-Марино или (Б) Лихтенштейн?

15. За сколько (примерно) дней Луна совершает полный оборот вокруг Земли: (А) 24 или (Б) 27?
16. В каком чае больше кофеина: (А) зеленом или (Б) черном?
17. Какая страна больше по площади: (А) Аргентина или (Б) Португалия?
18. Атомный вес какого элемента больше – (А) магния или (Б) цинка?
19. Какая страна больше по площади: (А) Аргентина или (Б) Казахстан?
20. Кто из животных быстрее бегает: (А) заяц или (Б) страус?
21. Какая страна больше по площади: (А) Лихтенштейн или (Б) Казахстан?
22. За что не дают Нобелевскую премию: (А) математику или (Б) медицину?
23. Что длиннее: (А) верста или (Б) вершок?
24. Кто прожил более долгую жизнь: (А) Петр I или (Б) Наполеон Бонапарт?
25. Кто из животных быстрее бегает: (А) лев или (Б) страус?
26. Какой из кинофестивалей старше: (А) Каннский или (Б) Венецианский?
27. Что длиннее: (А) вершок или (Б) дюйм?
28. На основе какой марки автомобиля была сделана первая модель «Жигулей»: (А) «Форд» или (Б) «Фиат»?
29. Какая страна больше по площади: (А) Германия или (Б) Украина?
30. Кто быстрее бегает: (А) заяц или (Б) человек?
31. В городе N два родильных дома: в одном рождается в среднем 10 детей в день (А), в другом 25 детей в день (Б). В среднем рождается одинаковое количество мальчиков и девочек. В каком из роддомов больше бывает таких дней в году, когда доля родившихся мальчиков превышает 60%?
32. Какое министерство было образовано раньше: (А) здравоохранения или (Б) путей сообщения?
33. Что длиннее: (А) сажень или (Б) ярд?
34. Кто из русских писателей похоронен в Тбилиси: (А) Грибоедов или (Б) Лермонтов?
35. Что длиннее: (А) верста или (Б) миля?
36. Какая из следующих стран не входит в Евросоюз: (А) Швейцария или (Б) Люксембург?
37. Окапи – это животное семейства: (А) сумчатых или (Б) жирафов?
38. Что длиннее: (А) сажень или (Б) верста?
39. Крокодил – это: (А) земноводное или (Б) пресмыкающееся?
40. Государство Бутан находится: (А) в Африке или (Б) в Азии?
41. В какой стране шире колея железной дороги: (А) в Финляндии или (Б) в Швеции?
42. Кто из животных быстрее: (А) гепард или (Б) страус?
43. Кто из животных быстрее бегает: (А) лев или (Б) гепард?
44. В каком городе численность населения раньше превысила 1 млн человек: (А) в Лондоне или (Б) в Багдаде?
45. До какого города расстояние от Москвы больше: (А) Орла или (Б) Санкт-Петербурга?

Prognosis Calibration in Binary Choice Tasks

Diana Kolesnikova
Research assistant, Laboratory for Experimental and Behavioral Economic ICEF NRU HSE
E-mail: dianakolesnikova@gmail.com

Alexis Belianin
Lecturer and Head, Laboratory for Experimental and Behavioural Economics, ICEF NRU HSE
E-mail: icef-research@hse.ru

Address: ICEF NRU HSE, 26 Shabolovka Str, Moscow, Russia 119049

Abstract

This works focuses on the problem of confidence in decisions made. Existing studies of prognosis calibration (defined as the ratio between the correct response frequency and the subjective evaluation of response correctness) have revealed that people tend to overestimate the probability of correct responses («overconfidence effect»). The authors hypothesized that subjective estimates of confidence in decisions made and the accuracy of prognosis calibration depend on the method used to measure confidence. An experimental study was performed in a sample of school students (N=50) and university students (N=36). The participants answered common knowledge questions and evaluated the degree of confidence in the correctness of their answers on verbal scales, using either a standard scale or a duplex scale (rating their confidence in both chosen and rejected answer options). Finally, the participants made bets that the chosen alternatives would be correct. The results indicated that the «overconfidence effect» was weaker in the duplex scale condition, confirming the hypothesis. We explain this finding by the notion of residual uncertainty that remains when two answer options are compared, but not in the standard scale setting, when the respondents' attention is concentrated only on the chosen answer option. The association between the level of confidence and prognosis quality had an inverse linear shape in the standard scale setting and a parabolic shape in a duplex scale setting. Finally, we found that in cases when the decisions were rational and the confidence ratings were consistent, bets predicted prognosis quality more reliably than verbal estimates did.

Keywords: Binary choice, prognosis calibration, Brier score, overconfidence, certitude, duplex scale, bet.

References

Allwood, C. M., & Granhag, P. A. (1996). Realism in confidence judgments as a function of working in dyads or alone. *Organizational Behavior and Human Decision Processes, 66*(3), 277–289.

Aukutsionek, S. P., & Belianin, A. V. (2001). Quality of forecasts and business performance: A survey study of Russian managers. *Journal of Economic Psychology, 22*(5), 661–692.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5 Suppl), 2–23.

Blavatskyy, P. R. (2008). *Betting on own knowledge: Experimental test of overconfidence*. University of Zurich Institute for Empirical Research in Economics W P 358.

Bornstein, B. H., & Zickafoose, D. J. (1999). I know I know it, I know I saw it: The stability of overconfidence across domains. *Journal of Experimental Psychology: Applied, 5*(1), 76–88.

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review, 89*(1), 306–318.

Cesarini, D., Sandewall, Ö. & Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior and Organization, 61*(3), 453–470.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 928–935.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and undercon dence: The role of error in judgment processes. *Psychological Review, 101*(3), 519–527.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26*(1), 32–53.

Foster, D. P., & Vohra, R. (1999). Regret in the on-line decision problem. *Games and Economic Behavior, 29*(1), 7–35.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond «heuristics and biases». *European review of social psychology, 2*(1), 83–115.

Glaser, M., & Weber, M. (2010). Overconfidence. In H. Kent Baker, & John R. Nofsinger (Eds.), *Behavioral finance: Investors, corporations, and markets* (pp. 241–258). New York: Wiley.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*(2), 226–246.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A «lazy» algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*(5), 563–607.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678–703.

Kalai, E., Lehrer, E., Smorodinsky, R. (1999). Calibrated forecasting and merging. *Games and Economic Behaviour, 29*(1), 151–169.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica, 77*(3), 217–273.

Kirchler, E., & Maciejovsky, B. (2002). Simultaneous over-and underconfidence: Evidence from experimental asset markets. *Journal of Risk and Uncertainty, 25*(1), 65–85.

Koellinger, P., Minniti, M., & Schade, C. (2007). «I think I can, I think I can»: Overconfidence and entrepreneurial behavior. *Journal of Economic Psychology, 28*(4), 502–527.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*(2), 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann and G. Zeeuw (Eds.), *Decision making and change in human affairs. Theory and decision library* (Vol. 16, pp. 275–324). Springer Netherlands.

Murphy, A. H., & Brown, B. G. (1985). A comparative evaluation of objective and subjective weather forecasts in the United States. In *Behavioral decision making* (pp. 329–359). New York: Springer

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79*(387), 489–500.

Mussweiler, T., & Posten, A. C. (2012). Relatively certain. Comparative thinking reduces uncertainty. *Cognition, 122*(2), 236–240.

Ronis, D.L., & Yates, J.F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40*(2), 193–218.

Suantak, L., Bolger, F. & Ferrell W. (1996). The «hard-easy effect» in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67*(2), 201–221.

Wickens, C. D. (1992). *Engineering Psychology and Human Performance* (2nd ed.), New York: Harper Collins.

Wright, G., & Ayton, P. (1988). Decision time, subjective probability, and task difficulty. *Memory and cognition, 16*(2), 176–185.

Wu, S. W., Johnson, J. E., & Sung M. C. (2008). Overconfidence in judgements: the evidence, the implications and the limitations. *The Journal of Prediction Markets, 2*(1), 73–90.

Yates, J. F. (1990). *Judgment and decision making*. New Jersey: Prentice-Hall.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic forecasts. *Journal of Forecasting, 4*(1), 61–73.

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: beyond general knowledge overconfidence? *Organizational Behaviour and Human Decision Processes, 74*(2), 89–117.