# ТЕОРИЯ, МЕТОДОЛОГИЯ И МЕТОДЫ

DOI: 10.14515/monitoring.2016.6.02

**Правильная ссылка на статью:**
Капуза А. В., Тюменева Ю. А. Надежность и структура шкалы социальной желательности TALIS: оценка в рамках современной теории тестирования // Мониторинг общественного мнения: Экономические и социальные перемены. 2016. № 6. С. 14—29.
**For citation:**
Kapuza A., Tyumeneva Yu. Reliability and Dimensionality of the TALIS scale of social desirability: evidence from the Item Response Theory. *Monitoring of Public Opinion: Economic and Social Changes.* 2016. № 6. P. 14—29.

**А. В. Капуза, Ю. А. Тюменева**

## НАДЕЖНОСТЬ И СТРУКТУРА ШКАЛЫ СОЦИАЛЬНОЙ ЖЕЛАТЕЛЬНОСТИ TALIS: ОЦЕНКА В РАМКАХ СОВРЕМЕННОЙ ТЕОРИИ ТЕСТИРОВАНИЯ

НАДЕЖНОСТЬ И СТРУКТУРА ШКАЛЫ СОЦИАЛЬНОЙ ЖЕЛАТЕЛЬНОСТИ TALIS: ОЦЕНКА В РАМКАХ СОВРЕМЕННОЙ ТЕОРИИ ТЕСТИРОВАНИЯ

*КАПУЗА Анастасия Васильевна — аспирантка I курса Национального исследовательского университета «Высшая школа экономики», Москва, Россия.*
*E-MAIL: a.v.kapuza@gmail.com*
*ORCID: 0000-0003-4982-5663*

*ТЮМЕНЕВА Юлия Алексеевна — кандидат психологических наук, старший научный сотрудник Института образования Национального исследовательского университета «Высшая школа экономики», Москва, Россия.*
*E-MAIL: jutu@yandex.ru*
*ORCID: 0000-0002-2381-917X*

**Аннотация.** Один из способов контролировать влияние социальных ожиданий на ответы респондентов — использовать шкалы социальной

RELIABILITY AND DIMENSIONALITY OF THE TALIS SCALE OF SOCIAL DESIRABILITY: EVIDENCE FROM THE ITEM RESPONSE THEORY

*Anastasia V. KAPUZA[1] — First-year Post-Graduate Student*
*E-MAIL: a.v.kapuza@gmail.com*
*ORCID: 0000-0003-4982-5663*

*Yulia A. TYUMENEVA[1] — Candidate of Psychological Sciences, Senior Researcher*
*E-MAIL: jutu@yandex.ru*
*ORCID: 0000-0002-2381-917X*

[1] National Research University Higher School of Economics, Moscow, Russia

**Abstract.** One way to take control over the effect of social desirability on respondent answers is to introduce social desirability scales into questionnaire.

желательности вместе с основными вопросами. Шкала социальной желательности, включенная для этой цели в международное сравнительное исследование Teaching and Learning International Survey (TALIS) и использованная на русскоязычной выборке учителей, не проходила кросс-культурной адаптации. Кроме того, этот инструмент основан на шкале социальной желательности Кроуна-Марлоу, психометрические характеристики которой оценивались до сих пор только в рамках классической теории тестирования и с неоднозначными результатами. Чтобы восполнить дефицит в понимании валидности шкалы социальной желательности TALIS в рамках современной теории тестирования были проанализированы данные, полученные на репрезентативной выборке российских учителей. Результаты показали приемлемую надежность, существенную одномерность шкалы и, наряду с этим, ряд серьезных проблем в ее функционировании. На основе полученных результатов, включая симулированные данные, предлагаются меры по повышению качества психометрических свойств шкалы. Делаются фундаментальные выводы о структуре конструкта социальной желательности.

**Ключевые слова:** социальная желательность, шкала Кроуна-Марлоу, современная теория тестирования (IRT), TALIS, факторная структура, надежность

Social desirability scale included into the Teaching and Learning International Survey (TALIS) and used for the Russian-speaking sample of teachers was not cross-culturally adapted. Besides that, this tool is based on the Marlowe-Crowne Scale where the psychometric characteristics are assessed only according to the Classical Test Theory and have ambiguous results. To fill in the gap in our knowledge of validity of the TALIS scale of social desirability, the authors conducted a psychometric analysis using the Item Response Theory. The results showed good reliability, considerable unidimensionality, though poor scale functioning. Based on the obtained results including simulated data, measures to improve the quality of psychometric characteristics of the scale are proposed by the authors. The key findings concerning the structure of the social desirability construct are made.

**Keywords:** social desirability, Marlowe-Crowne Scale, Item Response Theory, factor structure, reliability

## Постановка проблемы

Широкое распространение самоотчетных методов в международных исследованиях качества образования, таких как TIMSS, PISA, PIAAC, закономерно привело к вопросу об их валидности особенно в отношении тем, чувствительных к куль-

турным и социальным различиям [Rickett et al, 2006; Burris et al, 2003; Hebert et al, 1997; Мягков, 2007; Мягков, Журавлева, 2010; Левада, 2000]. В опросах учителей, к примеру, необходимость высказаться о качестве получаемой методической поддержки или подтвердить использование какого-то педагогического метода, неизбежно будет находиться под сильным влиянием социальных норм. В итоге опросные данные будут отражать не только суждения учителей относительно тех или иных событий, но и приемлемость последних в данном культурном окружении, и отделить одно от другого очень сложно. Проблема заключается не в том, что социальные ожидания влияют на ответы, а в том, что они влияют на них не в одинаковой степени в разных странах. Это снижает надежность данных об измеренном конструкте и, в конечном счете, вредит валидности основанных на этих данных сравнений.

Один из способов контролировать влияние социальных ожиданий на ответы респондентов — использовать шкалы социальной желательности (СЖ) вместе с основными вопросами. Так, в международном сравнительном исследовании качества подготовки учителей TALIS (Teaching and Learning International Survey, http://www.oecd.org/edu/school/talis.htm) в 2013 г. появилась шкала СЖ. По данным о выраженности СЖ в разных странах планировалось произвести статистическую коррекцию основных данных опроса. Иными словами, предназначение шкалы СЖ — не в том, чтобы сравнить уровень СЖ в разных странах, а в том, чтобы уменьшить ошибочные вариации по другим переменным.

Но прежде чем использовать СЖ для международных сравнений, необходимо убедиться, что шкала СЖ функционирует как надежный и валидный инструмент на уровне отдельных стран. Оценке надежности и валидности шкалы СЖ посвящено наше исследование.

**Измерение социальной желательности**

Конструкт СЖ определяется, как стремление респондента по каким-либо причинам давать такие ответы, которые, как ему кажется, выставят его в наиболее выгодном свете. В рамках этого определения была построена однофакторная шкала Эвардса (Edward's Social Desirability Scale (ESDS)) [Edwards, 1957]. Позднее была предложена двухкомпонентная модель СЖ, которая включила самообман (неосознаваемое искажение представления о себе) и управление впечатлением (сознательное манипулирование своим образом с целью представить себя в наиболее выгодном свете) [Wiggins, 1964; Sackeim, Gur, 1978; Paulhus 1984]. К настоящему времени разработаны трех- и четырехфакторные модели СЖ [Paulhus, John, 1998; Paulhus, Reid, 1991].

О психометрических характеристиках шкалы СЖ, использованной на русскоязычной выборке в исследовании TALIS, данные не публиковались. Поэтому в качестве аналога мы опишем психометрические характеристики и структуру шкалы Кроуна-Марлоу (MCSDS) [Crowne, Marlowe, 1960], чьей сокращенной версией является шкала СЖ TALIS [OECD, 2014]. Хотя при создании шкалы СЖ TALIS вопросы шкалы Кроуна-Марлоу были не только сокращены по количеству, но и содержательно адаптированы для школьной действительности, теоретически новая шкала должна функционировать аналогичным образом.

Шкала Кроуна-Марлоу показывает высокую надежность (от 0,7 до 0,8) [Crino et al, 1983; Longman et al, 1989], однако о структуре MCSDS данные противоречивы. Шкала создавалась СЖ как одномерный конструкт, и есть исследования, подтверждающие ее однофакторную структуру [Seol, 2007], но при этом некоторые утверждения полной формы могли слабо коррелировать с общим баллом [Strahan, Gerbasi, 1972].

Позднейшие исследования показали двухфакторную структуру шкалы [Paulhus, 1984; Barger, 2002; Leite, Beretvas, 2005; Crino et al, 1983; Loo, Thorpe, 2000]. В одном из вариантов двухфакторной структуры к первому фактору относились утверждения, отражающие негативные действия, а ко второму — желаемые, позитивные, и корреляция между факторами была 0,84 [Greenwald, Clausen, 1970], что говорит в пользу одного подлежащего конструкта, принимающего две формы: преуменьшение нежелательное поведения и преувеличение желательного [Zerbe, Paulhus, 1987].

На основе MCSDS разработано множество коротких форм, которые, как и полная форма шкалы, от исследования к исследованию показывают неустойчивые характеристики надежности и факторной структуры [Barger, 2002; Loo, Thorpe, 2000; Ramanaiah, Schill, Leung, 1977; Ханин, 1976].

Интересующая нас шкала СЖ TALIS показала двухфакторную структуру (см. рабочий отчет [Van de Vijver, He, 2014]) с факторами (1) позитивного управления впечатлением (например, «Я всегда внимательно слушаю учащихся») и негативного управления впечатлением (например, «Мне доводилось говорить такие вещи, которые оскорбляли чувства моих коллег или учащихся»). Содержательно эти два фактора идентичны тем, что были описаны для шкалы MCSDS Гринвальдом и Клаузеном [Greenwald, Clausen, 1970], хотя в шкале СЖ TALIS факторы коррелируют гораздо слабее, чем в MCSDS (0,50 и 0,84 соответственно). Тем не менее, поскольку и коэффициент, и содержание факторов оставляют сомнения в двухкомпонентной структуре СЖ, его можно рассматривать и как унитарный, принимающий две формы: преуменьшение нежелательного поведения и преувеличение желательного.

**Цель, ограничения и значимость исследования**

Наша цель — оценить психометрические характеристики и структуру короткой и адаптированной для учителей версии шкалы СЖ Кроуна-Марлоу, которая использовалась на русскоязычной выборке в исследовании TALIS в 2013 г. Обобщая данные предыдущих исследований шкалы СЖ Кроуна-Марлоу, можно заключить, что наилучшими психометрическими качествами обладает ее полная форма.

Так как каждая теория имеет свойственные ей ограничения, кратко рассмотрим теорию, в рамках которой оценивалась шкала. До сих пор структуру шкалы моделировали в рамках классической теории тестирования (КТТ), анализируя сырые тестовые баллы. Как известно, в рамках КТТ невозможно отделить тестовые показатели от «способностей» респондентов (например: [Hambleton, Swaminathan, 2013; Fan, 1998]). Поэтому в рамках КТТ мы не могли сказать, в какой мере трудность задания была характеристикой самого задания, а в какой — способностей респондента. Возможно, неустойчивость психометрических характеристик MCSDS

принимали за различия в выборках. Поэтому, работая в рамках КТТ, мы всегда должны допускать возможность того, что полученные результаты являются уникальными для конкретной выборки и их нельзя распространить на генеральную совокупность.

Конфирматорный факторный анализ (КФА), зачастую использовавшийся до сих пор для оценки структуры шкалы, также работает с сырыми тестовыми баллами. Это значит, что все выборочные смещения влияют на результаты факторного анализа. Кроме того, КФА основан на линейной измерительной модели и требует интервальной или дихотомической шкалы и нормальности распределения (например, [Bollen, 1989; Satorra, Bentler, 1990; Flora, Curran, 2004]. В случае со шкалой СЖ Кроуна-Марлоу часть этих допущений не проверялась, другие же не релевантны (так как MCSDS имеет дело с порядковыми данными).

Современная теория тестирования (Item response theory, IRT) позволяет дополнить проведенный в рамках КТТ анализ шкалы по нескольким аспектам. Кроме хорошо известных преимуществ — независимость трудности заданий от способности респондентов, возможность работать на распределении показателей разной формы и пр. (см., например [Embretson, Reise, 2000]) — IRT позволяет моделировать структуру шкалы, факторизуя не тестовые баллы, а стандартизированные остатки. А результаты факторизации остатков не зависят от выборочных характеристик, и сам метод свободен от допущения о линейных связях между переменными, что дает ему преимущество перед КФА. Помимо перечисленного, моделирование в рамках IRT позволяет оценить функционирование ответных категорий, чего в КТТ сделать невозможно. Ответными категориями называются варианты ответов, предложенные на выбор респонденту для указания степени его согласия с предложенным утверждением. Анализ функционирования ответных категорий позволяет оценить необходимость и достаточность того количества ответных категорий, которое использует шкала (а именно, семи категорий ответов), а также способность каждой ответной категории отражать соответствующий уровень выраженности измеряемого признака (черты).

Таким образом, наше исследование добавляет новое знание о конструкте СЖ и о структуре версии шкалы Кроуна-Марлоу, которая использовалась в TALIS.

Практическую значимость настоящей работы мы видим в проверке валидности шкалы на российской выборке. Частая критика шкалы Кроуна-Марлоу за противоречивые психометрические и структурные характеристики ставит под сомнение её использование в качестве основы для шкалы СЖ в международном исследовании (например, [Loo, Thorpe, 2000; Barger, 2002]). Тем более, что данные по СЖ предназначены для коррекции остальной части опроса, а использованная в TALIS шкала СЖ не проходила адаптации, необходимой для любого психологического инструментария при его переносе в другую культурную ситуацию.

## Метод

**Выборка.** Выборку составили 3972 респондентов-учителей, участвовавших в исследовании TALIS в 2013 г.: 3456 женщин (87 %), 516 мужчин (13 %); средний возраст 45,5 лет (диапазон от 19 до 76 лет, SD = 11,6 лет).

**Инструмент.** Шкала СЖ TALIS состоит из 10 утверждений с ликертовской шкалой ответов от 1 (полностью не согласен) до 7 (полностью согласен), с нейтральным вариантом «4» (Приложение 1).

**Последовательность анализа.** Анализ психометрических свойств шкалы СЖ выполнен в рамках Item Response Theory (IRT) в программе Winsteps [Linacre, 2011] в несколько этапов:

1. Анализ надежности.
2. Анализ размерности.
3. Выбор модели, подходящей для оценки функционирования категорий (partial credit model vs. rating scale model).
4. Выявление потенциально проблемных утверждений и нестандартных профилей ответов респондентов.
5. Сопоставление на одной шкале распределений уровня выраженности признака респондентов и трудности утверждений.
6. Анализ функционирования ответных категорий.
7. Анализ эффектов возможных перегруппировок ответных категорий.

## Результаты

1. *Анализ надежности*

Программа Winsteps сообщает надежность, эквивалентную классической надежности KR-20 или Альфе Кронбаха [Linacre, 2011]. Для шкалы СЖ TALIS надежность составила 0,67, что близко к наиболее низким коэффициентам надежности, сообщавшимся в исследованиях шкалы Кроуна-Марлоу (≈0,7). Однако этот показатель приемлем для психологических опросников и методик, имеющих небольшое количество заданий.

Все утверждения имеют удовлетворительную корреляцию с общим баллом по шкале от 0,30 до 0,45 при нижней приемлемой границе коэффициента 0,2; кроме того, утверждения шкалы имеют маленькую ошибку измерения (0,01—0,02). Учитывая согласованность с общим баллом по тесту и величину ошибки, надежность шкалы можно признать удовлетворительной.

2. *Анализ размерности*

Одномерность подразумевает преобладающее влияние одной латентной черты на выполнения всех заданий. Для предположительно одномерных шкал — это главнейшая характеристика, так как в противном случае тестовый балл не может интерпретироваться как оценка уровня выраженности измеряемого признака [McDonald, 1981; Карданова, 2016].

Для преодоления ограничений факторного анализа сырых баллов мы использовали факторный анализ стандартизированных остатков. Чтобы избежать влияния прочих факторов, которые могли влиять на ответы респондентов, факторный анализ стандартизированных остатков проведен на симулированных данных [Linacre, 2011]. На симулированных данных выделилось два фактора, объясняющих 5,4 % и 5,0 % дисперсии соответственно. В общем случае фактор признается значимым только в том случае, если он объясняет больше 5 % дисперсии, но поскольку шкала СЖ TALIS состоит из небольшого числа утверждений, мы пренебрегли строгостью требований и посчитали, что факторный анализ стандартизированных остатков подтвердил вывод об одномерности шкалы.

3. *Выбор модели, подходящей для оценки функционирования категорий (partial credit model vs. rating scale model)*

IRT включает семейство математических моделей, предназначенных для описания взаимосвязи между уровнем выраженности черты у респондента и выполнения им тестовых заданий [Vrieze, 2012]. Выбор правильной модели имеет решающее значение для получения всех преимуществ от использования IRT [Kang, Cohen, Sung, 2009]. Применение модели, плохо подходящей к данным, может приводить к неверной оценке параметров заданий и респондентов [Walker, Beretvas, 2003] или функционирования категорий [Bolt, 2002].

Наиболее часто для тестов с политомическими шкалами, какой является шкала СЖ, используются Rating Scale Model (RSM) [Andrich, 1978] и Partial Credit Model (PCM) [Masters, 1982]. При использовании модели RSM предполагается, что каждое утверждение сопровождается одинаковым количеством ответных категорий и трудности перехода между соседними категориями одинаковы. В свою очередь, при использовании модели PCM допускается различное количество категорий ответов и предполагается, что трудность перехода от категории к категории не обязательно возрастает с «порядковым номером» категории. Другими словами, в модели RSM переход к каждой следующей категории подразумевает повышение уровня выраженности признака испытуемого, а для модели PCM это не является обязательным. При использовании психологического опросника следует ожидать постепенного повышения трудности выбора категории по мере повышения ее порядкового номера. К примеру, для утверждения «Я всегда внимательно слушаю учащихся» выбрать категорию ответа «полностью согласен» должно быть труднее, чем «скорее согласен». Поэтому нам было важно удостовериться, что модель RSM больше подходит нашим данным, чем PSM.

Для оценки того, насколько хорошо модель RSM подходит данным, нами использовались статистики AIC [Akaike, 1974] и BIC [Schwarz, 1978]. На данный момент нет единого мнения о предпочтительности того или иного критерия, поэтому, как правило, для сравнения моделей используются оба: чем меньше их величины, тем лучше модель. Значения критериев, полученные в нашем анализе, говорят о том, что допущение об одинаковой трудности перехода к каждой следующей категории в различных утверждениях не подтверждается: RSM хуже подходит данным, чем PCM (табл. 1). Поэтому дальнейший анализ мы проводили в рамках модели PCM.

*Таблица 1.* ***Индексы AIC и BIC для RSM и PCM моделей***

| Модель | AIC | BIC |
|---|---|---|
| RSM | 93208 | 93308 |
| PCM | 90761 | 91110 |

4. *Выявление потенциально проблемных утверждений и нестандартных профилей ответов респондентов*

Важным преимуществом IRT по сравнению с КТТ является инвариантность оценок характеристик заданий относительно респондентов и инвариантность оценок мер респондентов относительно заданий. Другими словами, различные

выборки могут иметь разные распределения уровня подготовленности респондентов, но респонденты должны иметь равную вероятность ответить на задания одной и той же трудности независимо от принадлежности к выборке. Для этого отдельно исследуются согласие с моделью и утверждений, и респондентов. Согласие утверждений с моделью рассматривается для определения возможных проблем с формулировкой утверждений или их ответных категорий. Согласие данных респондента с моделью оценивается для выявления нестандартных профилей ответов респондентов, которые появляются, например, когда респондент отвечал путем случайного выбора ответа или угадывания [Reise, 1990].

Для исследования согласия с моделью утверждений и респондентов использовались соответствующие статистики согласия — INTFIT и OUTFIT MNSQ, — указывающие, насколько точно выбранная математическая модель предсказывает реальные ответы респондентов. В моделях Раша эти статистики принимают значения от 0 до +∞. Если данные полностью совпадают с предсказанными моделью, статистики принимают значение 1. Если данные имеют недостаточное согласие с моделью измерения, статистики принимают значения больше или меньше единицы; допустимыми являются значения статистик от 0,5 до 1,5 [Linacre, 2002b].

**Согласие утверждений с моделью.** Для шкалы СЖ среднее и стандартное отклонение INFIT статистики равны 1 и 0,2, для OUTFIT статистики 1,1 и 0,37 соответственно. Только для утверждения № 9 («Я признаюсь, что я чего-то не знаю, если учащийся задает вопрос в классе на уроке») обе статистики неудовлетворительны (табл. 2). Это утверждение является и самым трудным во всем тесте — с остальными утверждениями шкалы легко согласиться (т. е. они являются легкими). Таким образом, в целом утверждения шкалы функционируют удовлетворительно и пересмотра, с этой точки зрения, требует только задание № 9.

*Таблица 2.* ***Статистические данные по заданиям теста.***

| № | Оценка трудности | Ошибка измерения | Корреляция с баллом | MNSQ Infit | MNSQ Outfit |
|---|---|---|---|---|---|
| 1 | –0,19 | 0,02 | 0,41 | 0,87 | 0,78 |
| 2 | –0,08 | 0,02 | 0,45 | 0,98 | 1 |
| 3 | 0,21 | 0,01 | 0,42 | 1,11 | 1,23 |
| 4 | –0,08 | 0,02 | 0,39 | 0,98 | 0,98 |
| 5 | –0,24 | 0,02 | 0,35 | 0,84 | 0,81 |
| 6 | –0,08 | 0,01 | 0,4 | 0,87 | 0,85 |
| 7 | –0,18 | 0,02 | 0,37 | 0,87 | 0,96 |
| 8 | –0,14 | 0,02 | 0,37 | 1,00 | 1,02 |
| 9 | 1,03 | 0,01 | 0,41 | 1,54 | 2,1 |
| 10 | –0,24 | 0,02 | 0,3 | 0,9 | 1,27 |

**Согласие ответов респондентов с моделью.** Аналогично мы рассмотрели статистики согласия по респондентам для выявления профилей нестандартных ответов. Среднее и стандартное отклонение статистик равно 1,07 и 0,86 для INFIT и 1,08 и 1,27 для OUTFIT соответственно. При таком большом стандартном отклонении, следует признать, что данные большого количества респондентов в выборке не совпадают с предсказанными моделью. Это могло произойти из-за небольшого количество заданий в шкале, когда даже один или два неожиданных ответа могут значительно увеличить расхождение реальных данных с модельными [Карданова, 2016].

Поэтому для дальнейшего анализа психометрических свойств шкалы должны быть исключены респонденты, показавшие значительное расхождение реальных данных с модельными. При рассмотрении согласия респондентов с моделью оказалось, что в 845 случаях (22%) индивидуальная статистика MNSQ респондентов больше допустимого значения 1,5. Так как количество заданий небольшое, мы выбрали в качестве допустимого значение 2,0; оно было превышено в 529 случаях (14% респондентов). После удаления этих 529 случаев среднее и стандартное отклонение индивидуальных статистик составило 1,03 и 0,62 для INFIT и 0,94 и 0,67 для OUTFIT, соответственно. Таким образом, после удаления респондентов, показавших значительное расхождение реальных данных с модельными, общие статистики респондентов улучшились. В то же время самые общие характеристики выборки практически не изменились: средний возраст респондентов — 45,5 лет (стандартное отклонение 11,5), 87% респондентов — женщины. Дальнейший анализ был проведен на массиве без исключенных респондентов.

5. *Анализ распределений уровня выраженности признака респондентов и трудности утверждений, расположенных на одной шкале*

Карта переменных позволяет удостовериться, что уровень выраженности признака (в данном случае СЖ) респондентов и трудности заданий распределены одинаково, что позволяет свести ошибку измерения к минимуму [Карданова, 2016]. На рисунке 1 показано распределение трудности утверждений шкалы СЖ и уровня выраженности СЖ у респондентов относительно друг друга на общей метрической шкале логитов. Слева от шкалы расположены респонденты, справа — задания. На шкале буквами М обозначены средний уровень выраженности признака респондентов (слева от шкалы) и средняя трудность утверждений (справа от шкалы), которая также принимается за ноль. В верхней части карты находятся респонденты с наиболее выраженным уровнем черты и наиболее трудные задания. Выделим несколько важных моментов на карте. Во-первых, можно видеть, что большинство оценок респондентов, как и средняя их оценка, расположены выше средней трудности утверждений (выше нуля). Это говорит, во-первых, о высокой выраженности СЖ в нашей выборке. Во-вторых, шкала СЖ с большой погрешностью измеряет респондентов с высоким уровнем выраженности признака. В частности, семь из десяти утверждений шкалы СЖ практически бесполезны для оценивания таких респондентов из-за «потолочного эффекта», когда уровень СЖ у респондентов выше, чем могут «уловить» утверждения.

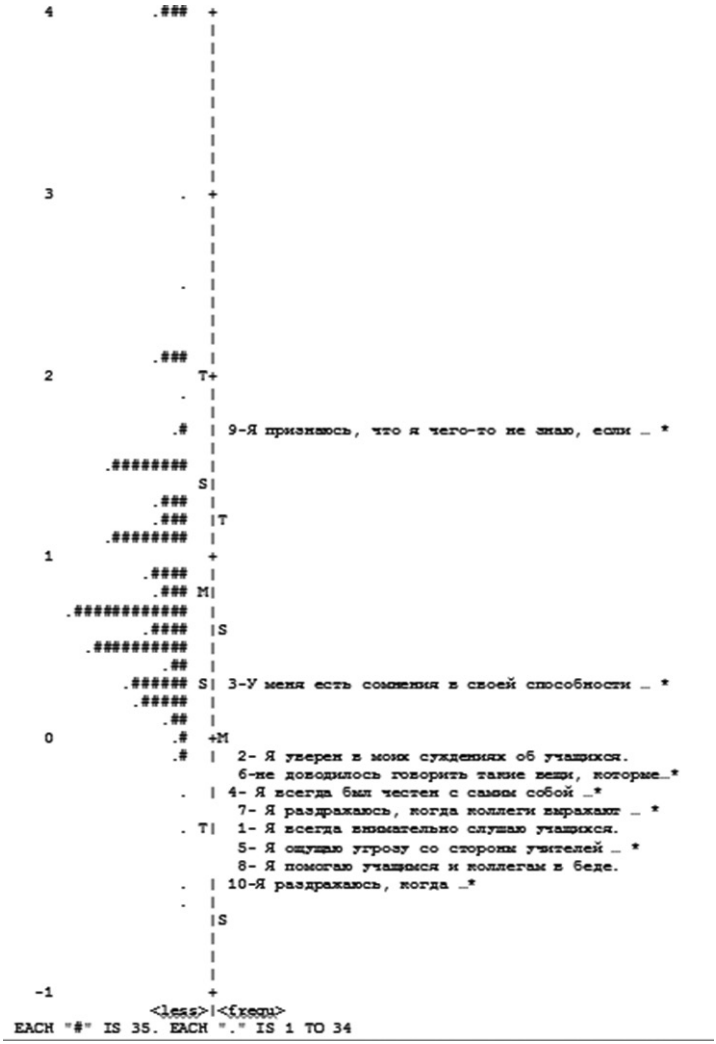


*Рисунок 1.* Карта переменных

***Примечание:*** На карте используются следующие обозначения: M — среднее значение, S и T — одно и два стандартных отклонений соответственно. Слева от шкалы расположены респонденты, справа — задания. В верхней части карты находятся наиболее подготовленные респонденты и наиболее трудные задания. Подготовленность в случае личностного опросника интерпретируется как уровень выраженности признака.

6. *Функционирование ответных категорий*

Наличие в личностных опросниках упорядоченных по возрастанию шкал ответов предполагает, что респонденты с большим уровнем исследуемого признака выбирают и более высокие значения ответной шкалы. Например, респондент, выбравший значение шесть на шкале Лайкерта, имеет более высокий уровень черты, чем респондент, выбравший на этой шкале значение четыре. В противном случае, значения категорий не несут никакой информации или у респондентов

возникают проблемы с определением смысла категорий. Это называется упорядоченностью категорий. Для проверки соответствия ответных категорий требованию упорядоченности рассматривается (1) средний уровень выраженности признаков у респондентов, выбравших эту категорию, и (2) пороговые оценки — уровень выраженности признака, при котором происходит переход от выбора категории с меньшим значением к выбору категории с большим значением (Linacre, 2002a). В шкале СЖ TALIS это требование частично выполнялось только в заданиях №№ 5 («Я ощущаю угрозу со стороны учителей, которые успешны в своей работе»), 6 («Мне доводилось говорить такие вещи, которые оскорбляли чувства моих коллег или учащихся»), 7 («Я раздражаюсь, когда коллеги выражают свои идеи, которые отличаются от моих») и 8 («Я помогаю учащимся и коллегам в беде»). В этих заданиях средний уровень респондентов возрастал вместе со значениями выбранной категории ответа. В остальных шести заданиях респонденты с одинаковым уровнем выраженности признака выбирали ответные категории неупорядоченно. Важно отметить, что в этих четырех заданиях упорядоченность категорий могла быть достигнута за счет их крайней легкости. В то же время второе требование — возрастание пороговых оценок — не выполнялось ни в одном задании.

Так как ни одно задание шкалы СЖ не показывает приемлемого функционирования ответных категорий, что говорит об их малой информативности, мы попробовали перегруппировать категории. Мы предполагаем, что сгруппировав исходные категории в более крупные, мы повысим их информативность. К примеру, если в 7-балльной шкале хорошо «работают» только крайние категории, а средние категории почти не различаются респондентами, то можно проверить создать вместо трех средних категорий одну, которая возможно будет хорошо функционировать.

7. *Перегруппировка категорий*

Как показал анализ, ни один из возможных вариантов перегруппировки категорий шкалы СЖ не дал значительных улучшений в функционировании ответных категорий. Наиболее удовлетворительной является перегруппировка категорий 1234567 в категории 1122233 соответственно, т. е. превращение семибалльной шкалы в трехбалльную. Улучшение заключается в том, что при такой перегруппировке наблюдается монотонное возрастание среднего уровня выраженности признака респондентов, выбирающих категорию ответа с более высоким «номером». Несмотря на улучшение качества функционирований категорий при подобной перегруппировке, у шкалы СЖ остается несколько проблем. Во-первых, вследствие уменьшения числа ответных категорий снижается дисперсия ответов, что, в свою очередь, приводит к снижению надежности шкалы до 0,61. Во-вторых, по-прежнему во всех утверждениях, кроме девятого, около 80 % респондентов выбирают третью, крайнюю, категорию. Таким образом, перегруппировка исходных категорий шкалы не решает всех проблем.

## Обсуждение

Обобщим результаты психометрического анализа шкалы СЖ TALIS. Качество этого инструмента отвечает двум важным требованиям: шкала имеет приемлемую надежность и существенно одномерную структуру. Последнее особенно важно, так

как на теоретическом уровне предшественница этой шкалы — шкала СЖ Кроуна-Марлоу — создана в рамках понимания СЖ как одномерного конструкта. Ранее при факторизации первичных баллов шкалы СЖ TALIS выделялись два фактора: позитивного и негативного управления впечатлением [Van de Vijver, He, 2014]. Содержательно эти факторы интерпретировались одинаково — как управление впечатлением, и скорее всего эта двухфакторная структура была обязана двунаправленности утверждений шкалы, так как факторы строго соответствовали направленности входящих в них прямых и обратных утверждений. Подобное выделение факторов часто встречается в шкалах с прямыми и обратными утверждениями, что является отдельной проблемой при определении структуры измеряемого конструкта (например, [Bagozzi, 1993; Marsh, 1996]).

Помимо этого, сомнения в полученном прежде двухфакторном решении основывались на свойствах конфирматорного факторного анализа, который предъявляет свои требования к выборке. В случае IRT моделирования мы были избавлены от влияний на факторную структуру особенностей выборки, так что с точки зрения конструктной валидности шкалы результаты проведенного исследования дают существенную поддержку пониманию СЖ как унитарного конструкта.

Основная проблема этой шкалы — функционирование ответных категорий. Прежде всего, данные показывают, что шкала не соответствует ожидаемой рейтинговой (RSM) модели: переход от одной ответной категории к последующей не связан с ростом выраженности СЖ у респондентов, что противоречит самой идее измерения психологических черт или состояний. Детальный анализ показал, что в большинстве утверждений функционируют только три категории: крайние степени согласия и несогласия и средняя категория. Это говорит о том, что 7-бальная шкала ответов избыточна. Более того, вероятность перехода от категории к категории неодинакова для различных утверждений, что говорит о том, что одни и те же категории имеют разный смысл для респондентов в различных утверждениях.

Техническая перегруппировка категорий не решает всех проблем с их функционированием, так что для улучшения психометрических свойств шкалы СЖ TALIS мы предлагаем предпринять следующие дополнительные меры. Во-первых, увеличить число утверждений, что позволит повысить надежность шкалы и возможно приведет к лучшему соответствию ее модели RSM, так как значимость отдельных отклонений в выборе ответных категорий респондентами снизится. Во-вторых, оптимизировать число категорий. В-третьих, возможно, одной из причин неудовлетворительного функционирования является их смысловая неопределенность. Для устранения этой причины было бы желательно назначить каждой категории более содержательные наименования, уменьшив при этом число самих категорий до четырех; в этом случае мы одновременно исключили бы «среднюю» категорию, которая сама по себе может являться показателем высокого уровня СЖ: как было отмечено выше, эту категорию часто выбирали респонденты с высоким уровнем СЖ.

С распространением личностных опросников в международных исследованиях проблема СЖ стала играть важную роль в оценке достоверности различий между странами. Поэтому статистический учет социальной желательности, в том числе и с помощью специальных шкал, становится все более востребованным направ-

лением исследований. Использование измерительных инструментов на международном уровне предполагает реализацию большого пласта психометрических исследований на национальном уровне. Кроме того, как мы показали, такие исследования могут давать и новые фундаментальные данные об оцениваемых психологических конструктах.

**Список литературы (References)**

*Карданова Е. Ю*. Моделирование и параметризация тестов: основы теории и приложения // Journal of European Social Policy. 2016. Т. 26. № . 3. С. 1—20. [Kardanova E. Yu. (2016) Modeling and test parameterization: theory foundations and applications. *Journal of European Social Policy*. Vol. 26. No. 3. P. 1—20] (in Russian).

*Левада Ю.* Человек лукавый: двоемыслие по-российски // Мониторинг общественного мнения : Экономические и социальные перемены. 2000. № 1. С. 19—27. [Levada Yu. (2000) Homo Prevaricate: Russian Doublethink. *Monitoring of Public Opinion: Economic and Social Changes*. No. 1. P. 19—27] (in Russian).

*Мягков А. Ю.* Искренность респондентов в сенситивных опросах: Методы диагностики и стимулирования. Иваново : Ивановский гос. энергетический ун-т им. В. И. Ленина, 2007. [Myagkov A. Yu. (2007) Iskrennost' respondentov v sensitivnykh oprosakh: Metody diagnostiki i stimulirovaniya [Respondents' sincerity in sensitive surveys: methods of diagnostics and stimulation]. Ivanovo: V. I. Lenin Ivanovo State Power University] (in Russian).

*Мягков А. Ю., Журавлева С. Ю.* О достоверности ответов респондентов в телефонном интервью // Социологические исследования. 2010. № 10. С. 81—93. [Myagkov A. Yu., Zhuravleva S. Yu. (2010) On reliability of the respondents answers in the telephone interviews. *Sociological studies*. No. 10. P. 81—93] (in Russian).

*Ханин Ю. Л.* Шкала Марлоу-Кроуна для исследования мотивации одобрения. Л. : НИИ ФК, 1976. [Khanin Yu. L. (1976) Shkala Marlou-Krouna dlya issledovaniya motivatsii odobreniya [The Marlow-Crowne Approval Motivation Scale]. L.: NII FK] (in Russian).

*Akaike H. A.* (1974) New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*. Vol. 19. No. 6. P. 716—723.

*Andrich D. A.* (1978) Rating Formulation for Ordered Response Categories. *Psychometrika*. Vol. 43. No. 4. P. 561—573.

*Bagozzi R. P.* (1993) An Examination of the Psychometric Properties of Measures of Negative Affect in the PANAS-X Scales. *Journal of Personality and Social Psychology*. Vol. 65. P. 836—851.

*Barger S. D.* (2002) The Marlowe-Crowne Affair: Short Forms, Psychometric Structure, and Social Desirability. *Journal of personality assessment*. Vol. 79. No. 2. P. 286—305.

*Bollen K. A.* (1989) Structural Equations with Latent Variables. New York: John Wiley & Sons.

*Bolt D. M.* (2002) A Monte Carlo Comparison of Parametric and Nonparametric Polytomous DIF Detection Methods. *Applied Measurement in Education*. Vol. 15. No. 2. P. 113—141.

*Burris J. E., Johnson T. P., O'Rourke D. P.* Validating Self-Reports of Socially Desirable Behaviors.

*American* Association for Public Opinion Research-Section on Survey Research Methods. 2003. P. 32—36. Retrieved from URL: http://ww2.amstat.org/sections/srms/Proceedings/y2003/Files/JSM2003-000914.pdf.

*Crino M. D. et al.* (1983) Data on the Marlowe-Crowne and Edwards social Desirability Scales. *Psychological Reports*. Vol. 53. No. 3. P. 963—968.

*Crowne D. P., Marlowe D.* (1960) A New Scale of Social Desirability Independent of Psychopathology. *Journal of Consulting Psychology*. Vol. 24. No. 4. P. 349—354.

*Edwards A. L.* (1957) The Social Desirability Variable in Personality Assessment and Research. New York: Dryden.

*Embretson S. E., Reise S. P.* (2000) Item Response Theory for Psychologists. Hillsdale, NJ: Lawrence Erlbaum Associates.

*Fan X.* (1998) Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics. *Educational and Psychological Measurement*. Vol. 58. No. 3. P. 357—381.

*Flora D. B., Curran P. J.* (2004) An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data. *Psychological methods*. Vol. 9. No. 4. P. 466—491.

*Greenwald H. J., Clausen J. D.* (1970) Test of Relationship Between Yeasaying and Social Desirability. *Psychological Reports*. Vol. 27. No. 1. P. 139—141.

*Hambleton R. K., Swaminathan H.* (2013) Item Response Theory: Principles and Applications. N. Y.: Springer Science & Business Media.

*Hebert J. R. et al.* (1997) Gender Differences in Social Desirability and Social Approval Bias in Dietary Self-Report. *American Journal of Epidemiology*. Vol. 146. No. 12. P. 1046—1055.

*Kang T., Cohen A. S., Sung H. J.* (2009) Model Selection Indices for Polytomous Items. *Applied Psychological Measurement*. Vol. 33. No. 7. P. 499—518.

*Leite W. L., Beretvas S. N.* (2005) Validation of Scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement*. Vol. 65. No. 1. P. 140—154.

*Linacre J. M. et al.* (2002a) Optimizing Rating Scale Category Effectiveness. J Appl Meas. Vol. 3. No. 1. P. 85—106.

*Linacre J. M.* (2002b) What Do Infit and Outfit, Mean-Square and Standardized Mean. *Rasch Measurement Transactions*. Vol. 16. No. 2. P. 878.

*Linacre J. M.* A user's guide to WINSTEPS Program manual 3.71.0. 2011. Retrieved from http://www.winsteps.com/manuals.shtml.

*Longman R. S. et al.* (1989) A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate behavioral research*. Vol. 24. No. 1. P. 59—69.

*Loo R., Thorpe K.* (2000) Confirmatory factor analyses of the full and short versions of the Marlowe-Crowne Social Desirability Scale. *The Journal of social psychology*. Vol. 140. No. 5. P. 628—635.

*Marsh H. W.* (1996) Positive and Negative Global Self-Esteem: A Substantively Meaningful Distinction or Artifactors? *Journal of Personality and Social Psychology*. Vol. 70. No. 4. P. 810—819.

*Masters G. N.* (1982) A Rasch Model for Partial Credit Scoring. *Psychometrika*. Vol. 47. No. 2. P. 149—174.

*McDonald R. P.* (1981) The dimensionality of tests and items. *British Journal of mathematical and statistical Psychology.* Vol. 34. No. 1. P. 100—117.

*OECD.* TALIS 2013: Technical Report. 2014. http://www.oecd.org/edu/school/talis-technicalreport-2013.pdf.

*Paulhus D. L., John O. P.* (1998) Egoistic and moralistic biases in self-perception: the interplay of self-deceptive styles with basic traits and motives. *Journal of personality*. Vol. 66. No. 6. P. 1025—1060.

*Paulhus D. L., Reid D. B.* (1991) Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*. Vol. 60. No. 2. P. 307—317.

*Paulhus D. L.* (1984) Two-component models of socially desirable responding. *Journal of personality and social psychology*. Vol. 46. No. 3. P. 598—609.

*Ramanaiah N. V., Schill T., Leung L. S.* (1977) A test of the hypothesis about the two-dimensional nature of the Marlowe-Crowne Social Desirability Scale. *Journal of Research in Personality.* Vol. 11. No. 2. P. 251—259.

*Reise S. P.* (1990) A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*. Vol. 14. No. 2. P. 127—137.

*Rickett B., Orbell S., Sheeran P.* (2006) Social-cognitive determinants of hoist usage among health care workers. *Journal of Occupational Health Psychology*. Vol. 11. No. 2. P. 182—196.

*Sackeim H. A., Gur R. C.* (1978) Self-deception, self-confrontation, and consciousness. *Consciousness and self-regulation: Advances in Research*. N. Y.: Plenum. P. 139—197.

*Satorra A., Bentler P. M.* (1990) Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*. Vol. 10. No. 3. P. 235—249.

*Schwarz G. et al.* (1978) Estimating the dimension of a model. *The Annals of Statistics*. Vol. 6. No. 2. P. 461—464.

*Seol H.* (2007) A psychometric investigation of the Marlowe-Crowne Social Desirability Scale using Rasch measurement. *Measurement and evaluation in counseling and development*. Vol. 40. No. 3. P. 155—168.

*Strahan R., Gerbasi K. C.* (1972) Short, homogeneous versions of the Marlowe-Crowne Social Desirability Scale. *Journal of clinical psychology*. Vol. 28. P. 191—193.

*Van de Vijver F. J.R., He J.* (2014) Report on social desirability, midpoint and extreme responding in TALIS 2013. *OECD Education Working Papers*. No. 107.

*Vrieze S. I.* (2012) Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*. Vol. 17. No. 2. P. 228—243.

*Walker C. M., Beretvas S. N.* (2003) Comparing Multidimensional and Unidimensional Proficiency Classifications: Multidimensional IRT as a Diagnostic Aid. *Journal of Educational Measurement*. Vol. 40. No. 3. P. 255—275.

*Wiggins J. S.* (1964) Convergences among stylistic response measures from objective personality tests. *Educational and Psychological Measurement*. Vol. 24. P. 551—562.

*Zerbe W. J., Paulhus D. L.* (1987) Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review*. Vol. 12. No. 2. P. 250—264.

## ПРИЛОЖЕНИЕ

Русскоязычная версия шкалы СЖ TALIS

В какой мере Вы согласны или не согласны со следующими утверждениями, касающиеся Вашего личного отношения?

В каждом пункте выберите один вариант ответа.

| | Полностью не согласен | | … | Средне | … | | Полностью согласен |
|---|---|---|---|---|---|---|---|
| 1. Я всегда внимательно слушаю учащихся. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2. Я уверен в моих суждениях об учащихся. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 3. У меня есть сомнения в своей способности добиться успеха в качестве преподавателя. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 4. Я всегда был честен с самим собой по поводу своих преподавательских навыков. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 5. Я ощущают угрозу со стороны учителей, которые успешны в своей работе. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 6. Мне доводилось говорить такие вещи, которые оскорбляли чувства моих коллег или учащихся. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 7. Я раздражаюсь, когда коллеги выражают свои идеи, которые отличаются от моих. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8. Я помогаю учащимся и коллегам в беде. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 9. Я признаюсь, что я чего-то не знаю, если учащийся задает вопрос в классе на уроке. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 10. Я раздражаюсь, когда учащиеся просят меня о помощи. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |