Возможности построения шкал и проведения кросс-культурных сравнений с помощью SEM и IRT

Презентация: Евгений Н. Осин,
дискуссия: Елена Ю. Карданова,
НИУ ВШЭ

Семинар ИО НИУ ВШЭ, 29.11.2013

План доклада

- Модель и процедуры мультигруппового конфирматорного факторного анализа
- Сопоставление моделей IRT и SEM
- Пример использования мультигруппового КФА для установления эквивалентности измерительного инструмента

Модель КФА и процедура мультигруппового КФА

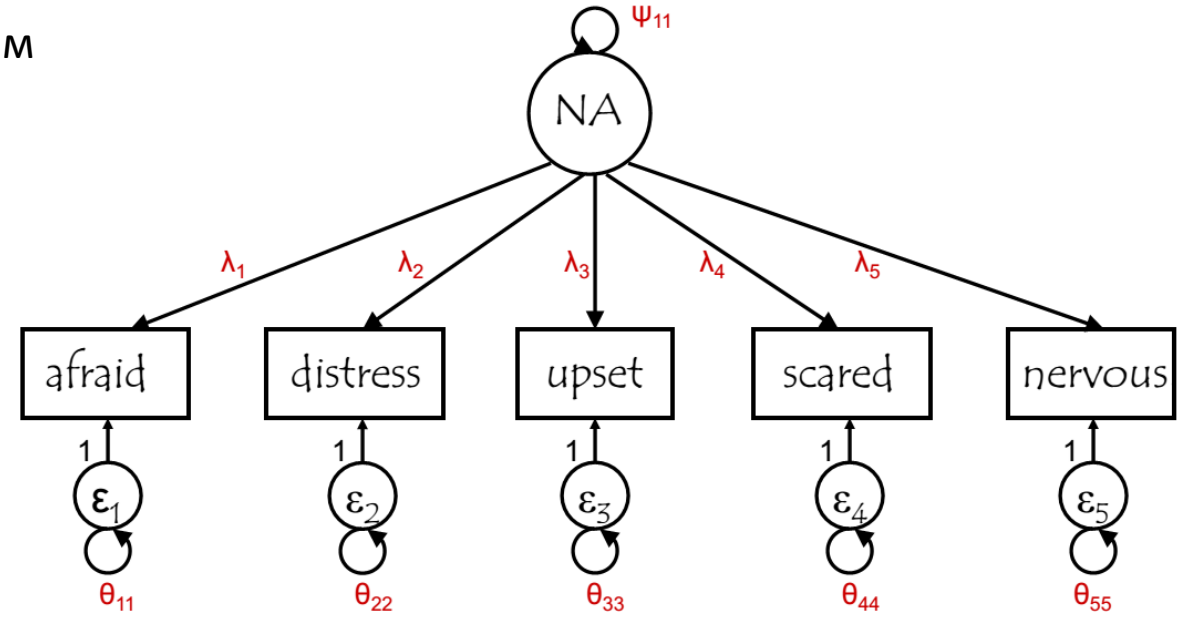
Типичные задачи КФА

- Проверка гипотез о структуре связей переменных
- Исследование психометрической эквивалентности инструментов в различных группах респондентов:
  - Версии опросника на разных языках;
  - Представители разных социальных групп.

Модель КФА

- Как и в случае IRT, мы предполагаем наличие латентных факторов, которые стоят за дисперсией пунктов.
- Формулировка априорных гипотез – на основе теоретических предположений или результатов эксплораторных методов (факторный анализ и пр.)

Красным цветом выделены оцениваемые (свободные) параметры



Модель КФА

$$\Sigma = \Lambda\Psi\Lambda^{T} + \Theta$$

$$\mu = \nu + \Lambda\alpha$$

- Σ = матрица дисп./ковариаций наблюдаемых переменных;
- Λ = матрица нагрузок наблюдаемых переменных на латентные факторы;
- Ψ = матрица дисперсий/ковариаций латентных факторов;
- Θ = матрица дисперсий/ковариаций остатков, или ошибок измерения наблюдаемых переменных (residuals);
- μ = вектор средних наблюдаемых переменных;
- ν = вектор остаточных средних (intercepts) наблюдаемых переменных;
- α = вектор средних латентных факторов.

Модель КФА

$$\Sigma = \Lambda\Psi\Lambda^{T} + \Theta \qquad \mu = \nu + \Lambda\alpha$$

Значения минимально необходимой части параметров фиксируются для идентификации модели:

- для идентификации шкалы латентного фактора – одна из нагрузок переменных на него либо его дисперсия;
- средние для латентных факторов фиксируются в одной группе (reference group).

Остальные параметры модели оцениваются: алгоритм минимизирует несоответствие между ожидаемой матрицей ковариаций (S) и воспроизведённой на основе параметров модели (Σ).

Алгоритм далеко не всегда сходится.

Модель КФА

$$\Sigma = \Lambda\Psi\Lambda^{T} + \Theta \qquad \mu = v + \Lambda\alpha$$

- 2 типа анализа:
  COVS (covariance structure) – только левое уравнение;
  MACS (mean and covariance structure) – со средними.
- Показатель соответствия модели – статистика χ^2, которая рассчитывается через сопоставление воспроизведённой матрицы ковариаций Σ с ожидаемой S.
- Считаем различие в вероятностных функциях нашей модели ($\log L_0$) и насыщенной ($\log L_1$), для которой Σ = S:

$$\log L = c - \frac{N}{2}\log|\Sigma| - \frac{N}{2}tr\left(S\Sigma^{-1}\right) \qquad \chi^2 = -2\log L_0 + 2\log L_1$$

Модель КФА

$$\Sigma = \Lambda \Psi \Lambda^{T} + \Theta$$

$$\mu = v + \Lambda \alpha$$

- 2 типа анализа:
  COVS (covariance structure) – только левое уравнение;
  MACS (mean and covariance structure) – со средними.
- В случае MACS в уравнении вероятностной функции прибавляется вектор m (ожидаемые средние):

$$\log L = c - \frac{N}{2}\log|\Sigma| - \frac{N}{2}tr\left(S\Sigma^{-1}\right) - \frac{N}{2}(m-\mu)^{T}\Sigma^{-1}(m-\mu)$$

$$\chi^2 = -2\log L_0 + 2\log L_1$$

Показатели соответствия в КФА

- Хи-квадрат: надёжный, но зависит от объёма выборки
- RMSEA: $d = \chi^2 - df\ /\ (N-1).\ \ \mathrm{RMSEA} = \mathrm{SQRT}[d\ /\ df]$
- CFI:

$$\mathrm{CFI} = 1 - [(\chi^2_H - \mathrm{df}_H)/(\chi^2_B - \mathrm{df}_B)]$$

where H = the hypothesized model, and B = the baseline model.

- Обычно показателями отличного соответствия модели данным считаются значения CFI>0,95 и RMSEA<0,05, приемлемого соответствия – CFI>0,9 и RMSEA<0,08.
- Но, как показывают данные симуляций, оптимальные значения CFI и RMSEA зависят также от числа переменных и латентных факторов в модели.

Индексы модификации в КФА

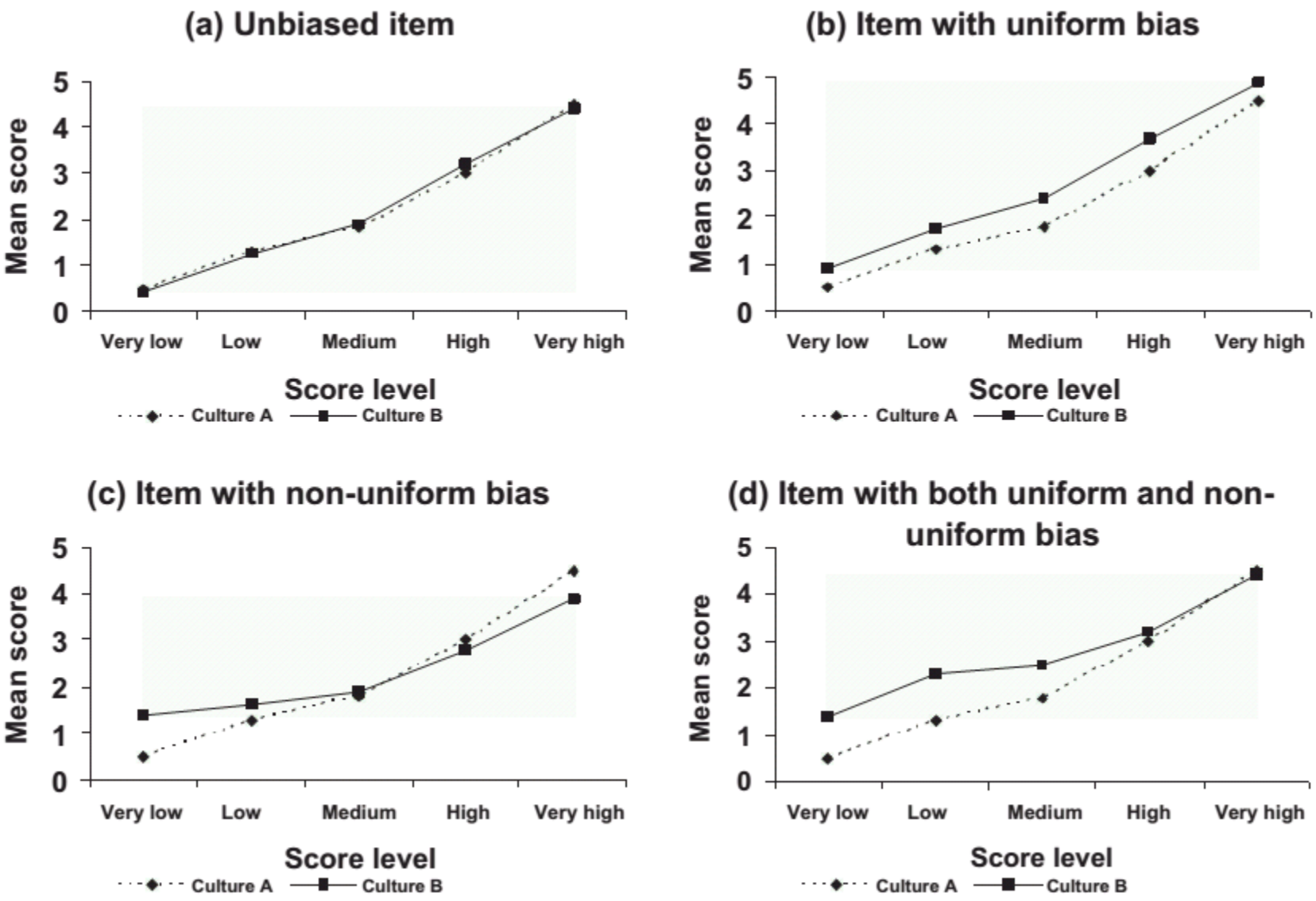
- Индексы модификации показывают, значимо ли (и насколько сильно) улучшится модель (уменьшится значение статистики хи-квадрат), если добавить новый свободный параметр:
  - ковариацию ошибок измерения переменных,
  - нагрузку переменной на дополнительный фактор,
  - снять ограничение на равенство значения параметра в отдельной выборке значениям аналогичных параметров в других выборках.
- Перебираются и оцениваются все возможные варианты добавления параметров.
- Чем больше параметров в модели, тем лучше она соответствует данным, но хуже обобщается на другие выборки.

Процедура мультигруппового КФА

1. На основе данных эксплораторного факторного анализа или теоретических предположений строится модель измерения для каждой из групп в отдельности:
   - Модель может быть доработана путём введения параметров на основе индексов модификации.
2. Среди параметров выделяются общие (равенство которых будет проверяться) и специфичные для конкретных групп (при наличии / необходимости).
3. Проверяется единая мультигрупповая модель, в которую вводятся ограничения на равенство значений параметров между группами:
   - Нагрузок переменных на факторы
   - Остаточных средних (intercepts) переменных
   - При необходимости: дисперсий ошибки измерения переменных, ковариаций латентных факторов, ковариаций ошибок измер-я…

Неэквивалентность параметров (DIF)

- **Неэквивалентность остаточных средних** (intercepts) = uniform bias → средние в группах будут различаться, но не из-за различий в измеряемом конструкте, а из-за особенностей формулировки пункта
- **Неэквивалентность факторной нагрузки** = non-uniform bias → различие средних в группах будет зависеть от степени выраженности измеряемого конструкта в группах. Этот вид bias обычно встречается реже и менее опасен.
- **Если мы сравниваем сырые баллы без учёта bias, мы рискуем сделать ложные выводы о значимых различиях между группами по измеряемому свойству.**



Примеры DIF (van de Vijver & Leung, 2011)

Неэквивалентность параметров (DIF)

- Только при эквивалентности нагрузок и остатков сравнение сырых баллов между группами имеет смысл:
- На практике, даже выполнение требования эквивалентности нагрузок и остаточных средних выполняется редко
- → можно установить **частичную эквивалентность** конструктов (Byrne, Shavelson, Muthen, 1989), выделив подмножество параметров, достаточное для осмысленного сопоставления групп по значениям латентных факторов (как минимум 1-2 пункта на шкалу)

Некоторые проблемы частичной MI

- Риск “capitalizing on chance” (→ могут быть проблемы с кросс-валидизацией), на малых выборках оценки параметров неустойчивы, индексы модификации менее информативны
- Параметры не являются независимыми → чем больше неэквивалентных параметров, тем сильнее проблемы с точностью оценки инвариантных параметров
- Свободные параметры, введённые в одной группе, отсутствуют в остальных → имплицитное допущение, что там они = 0.

CFA vs IRT: в чём разница?

CFA vs IRT (Brown, 2006)

- 1-факторный КФА аналогичен 2-PL модели IRT (для политомических переменных – Samejima's Graded Response model):
  - параметры сложности пунктов аналогичны остаточным средним пунктов (порогам);
  - параметры дискриминативности пунктов аналогичны факторным нагрузкам в КФА;
  - возможен даже взаимый перевод значений параметров моделей КФА ←→ IRT (Muthen, Kao, Burstein, 1991).
- Аналог 1-PL модели IRT – 1-факторный КФА с фиксированными нагрузками переменных на фактор.
- КФА-аналога для 3 PL модели пока нет.

CFA vs IRT (Brown, 2006)

- Аналог DIF-анализа – более гибкие мультигрупповые MIMIC-модели в КФА, их достоинства (Muthén):
  - ковариаты могут быть непрерывными и категориальными;
  - ковариаты могут иметь эффект на латентный фактор;
  - модели могут быть многомерными;
  - ошибки пунктов могут коррелировать.
- Различные способы оценки моделей: ML, WLS, Bayesian → разные виды распределений.

Сходства мультигрупповых CFA и IRT
(Raju, Lafitte, Byrne, 2002)

1. Оба подхода связывают неизмеряемый (латентный) конструкт с набором измеренных переменных.
2. Оба подхода позволяют оценить, будут ли одинаковыми истинные баллы (без учёта случайной погрешности) по пунктам/шкалам для людей из разных групп с одинаковыми уровнями неизмеренной черты.
3. Оба подхода не предполагают равенства параметров распределения неизмеренной черты в группах (impact).
4. Оба подхода позволяют оценить выраженность DIF и выявить источники проблемы.
5. Оба подхода позволяют строить Item Response Functions.

Различия мультигрупповых CFA и IRT

(Raju, Lafitte, Byrne, 2002)

1. **Связь латентного конструкта с истинными баллами по пунктам / субшкалам в IRT нелинейна, в CFA линейна (за исключением CFA для порядковых переменных).**
2. Для дихотомических пунктов более адекватна модель логистической регрессии в IRT, но для политомических пунктов характеристики линейной модели сопоставимы.
3. **В CFA хорошо разработана методология работы с несколькими (= многомерными) конструктами в нескольких группах. В IRT лучше разработаны одномерные модели.**
4. В CFA обсуждается необходимость равенства дисперсий ошибки пунктов, в IRT – стандартная ошибка.
5. IRT даёт информацию о вероятности выбора альтернатив в зависимости от уровня латентной черты.
6. **IRT позволяет оценить компенсаторный (суммарный) DIF для всей шкалы, CFA – только DIF для пунктов. Но CFA позволяет моделировать частично эквивалентные инструменты.**

Пример с данными NorBA

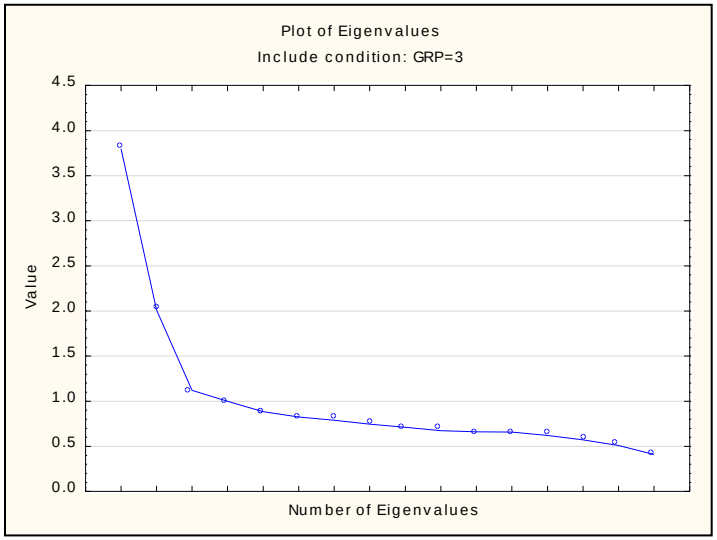
NorBA

- Международный опрос установок учителей
- Анкеты переведены на национальный язык с английского
- Для целей доклада мы взяли 1 блок анкеты:
  - D: установки по отношению к преподаванию.
- Использованы данные из 3 стран:
  - Россия (N=343),
  - Латвия (N=390),
  - Эстония (N=332).

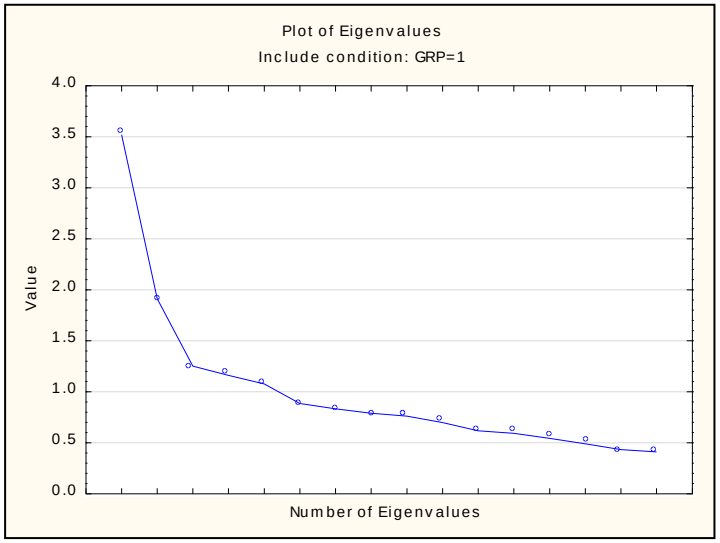
Часть D Анкеты: установки учителей

1. Проблемы повседневной и будущей жизни учеников являются значимым условием для развития их знаний
2. Обучение нужно основывать на заданиях с ясными правильными ответами и на основании тех идей, которые большинство учеников могут быстро усвоить
3. Объем усвояемого материала зависит от существующего на данный момент объема знаний учеников – поэтому так важно преподавать факты
4. Хорошие учителя показывают, как правильно решать задание
5. Роль учителя – способствовать исследовательской деятельности учеников
6. Ученики учатся лучше всего тогда, когда самостоятельно находят решения заданий
7. Ученикам нужно дать возможность самим поработать над практическими заданиями до того, как учитель покажет правильное решение
8. Учителя должны направлять учеников к их личным открытиям
9. Чтобы развивать концептуальное понимание у учеников, учителям необходимо использовать различные методы (соответствующие ситуации)
10. Учеников следует вовлекать в работу в небольших группах, где они могут объяснить свои новые идеи и выслушать идеи других учеников
11. Процессы мышления и рассуждения важнее, чем содержание конкретной учебной программы
12. Большинство видов деятельности требует использования имеющихся знаний и навыков по-новому
13. Учителю следует акцентировать внимание на использовании знаний и умений, приобретенных на других уроках, для решения заданий и понимания проблем
14. Ученики вместе со своими учителями разрабатывают критерии оценивания и/или средства оценивания
15. Оцениваться должны и практические задачи, проекты, исследования
16. Чтобы учебный процесс был эффективным, в классе должна быть тишина

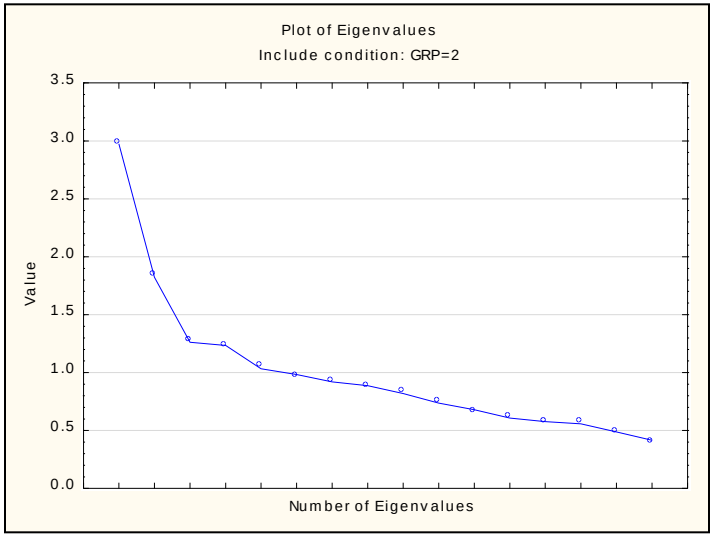
Двухфакторная структура? (МГК)



Россия



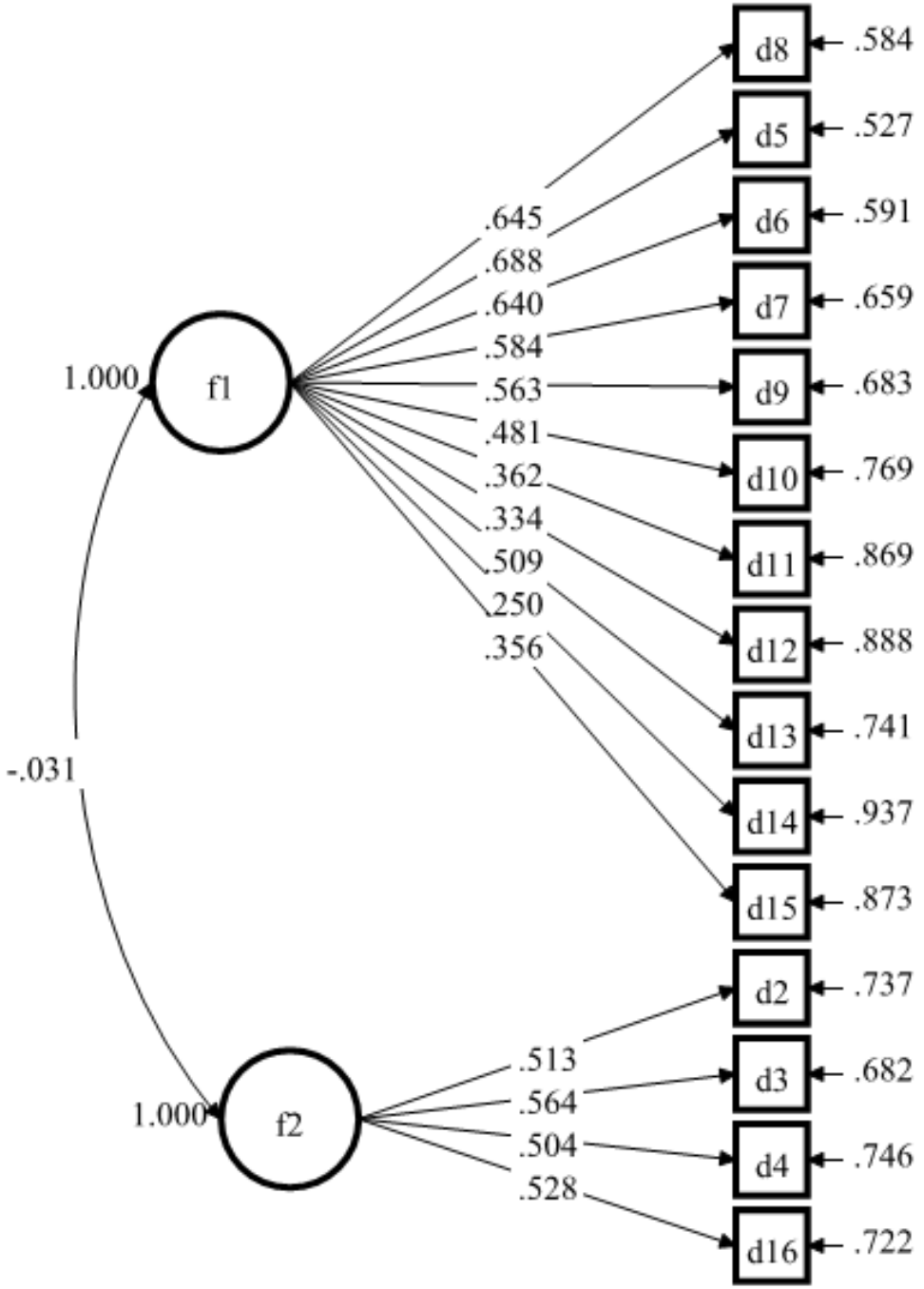
Латвия



Эстония

Теоретическая модель (Россия, станд. оценки параметров)

F1 = Конструктивизм,
F2 = Традиционализм



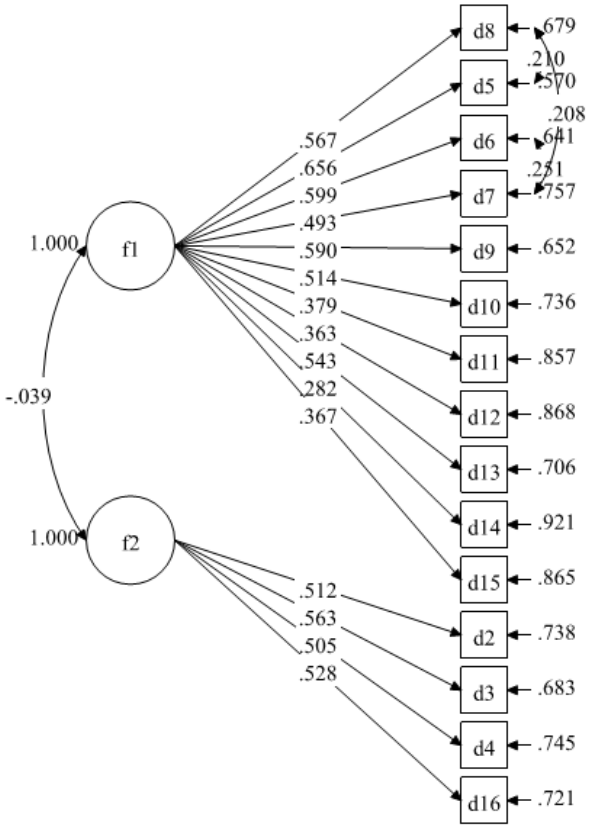
Метод анализа

- Неоднородность структуры:
  - в России – 2 чётких фактора, в Эстонии – 2 чётких и 2 слабых, плохо интерпретируемых фактора, в Латвии – нечто среднее между Россией и Эстонией;
  - решение: брать российскую модель, 2 фактора которой воспроизводятся везде, за основу, подфакторы моделировать корреляциями остатков.
- Есть пропущенные данные → FIML алгоритм MLR.
- Модели для порядковых переменных (WLSMV) работают лучше, чем модели для категориальных, но много переменных с «эффектами потолка» → не все категории представлены во всех группах.

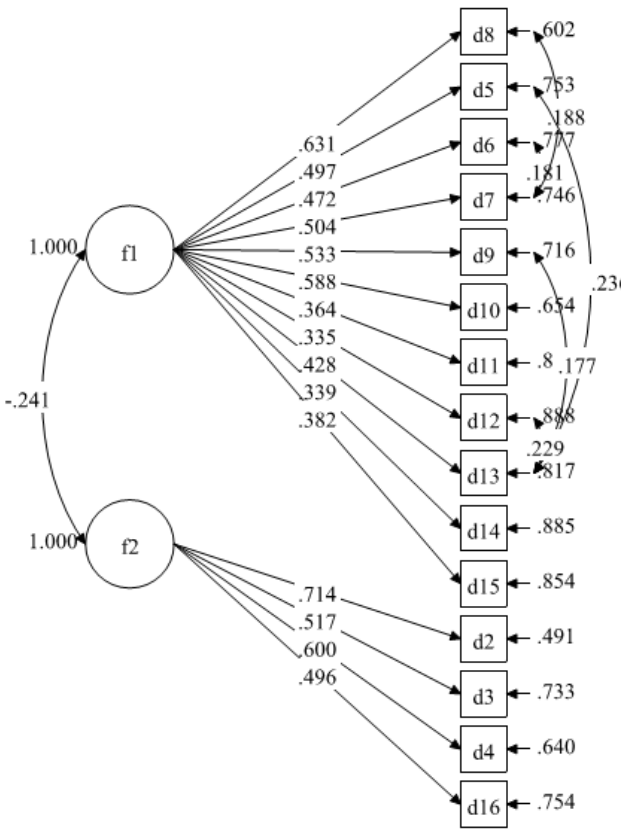
Введённые свободные параметры (независимая разработка 3 моделей)

	Россия (N=343)	**Латвия (N=390)**	**Эстония (N=332)**
Показатели теоретической модели измерения	X2(89)=144,97 CFI=0,919 RMSEA=0,043 (0,030-0,055)	X2(89)=228,24 CFI=0,834 RMSEA=0,063 (0,053-0,074)	X2(89)=202,71 CFI=0,763 RMSEA=0,062 (0,051-0,073)
Ковариации ошибок (в пределах фактора)	**6 и 7**, **7 и 8**, 5 и 8	**6 и 7**, **7 и 8**, 5 и 13, 9 и 13, 12 и 13	**6 и 7, 7 и 8**, 6 и 10, 11 и 14
Двойные нагрузки	---	---	Пункт 12 на F2
Показатели полученной модели измерения	X2(86)=112,45 CFI=0,962 RMSEA=0,030 (0,010-0,044)	X2(84)=167,65 CFI=0,900 RMSEA=0,051 (0,039-0,062)	X2(84)=140,15 CFI=0,883 RMSEA=0,045 (0,031-0,058)

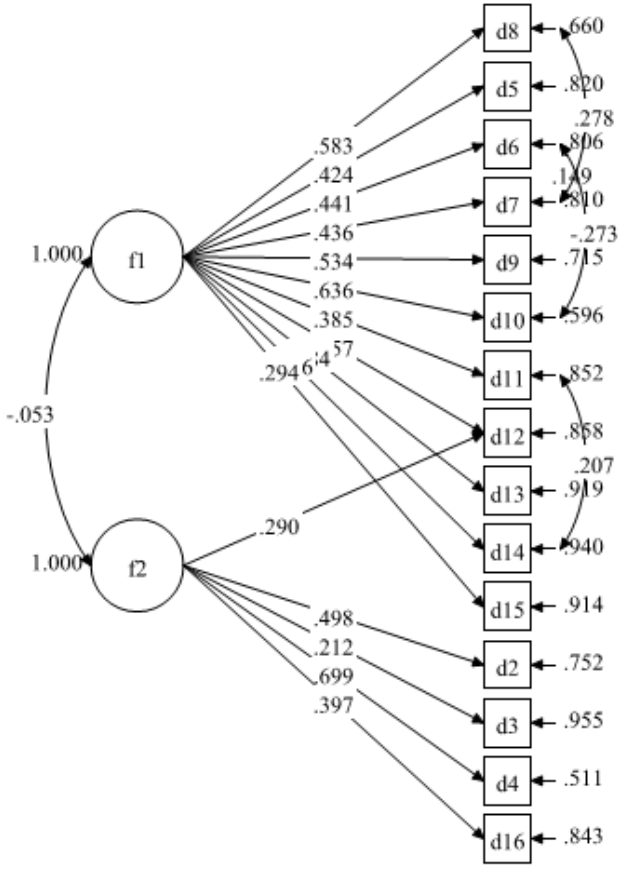
Итоговые модели измерения



Россия



Латвия



Эстония

Введённые свободные параметры (разработка мультигрупповой модели)

		Модель для 3 групп (N=1065)		
		Россия	Латвия	Эстония
1. Модель измерения		**X2(254)=421,89; CFI=0,917; RMSEA=0,043 (0,036-0,050)**		
2. Равенство нагрузок		X2(280)=478,17; CFI=0,901; RMSEA=0,045 (0,038-0,051)		
	Неэквивалентные	5[X2=13]	----	13[X2=13]
	Частичная эквивал.	**X2(278)=452,28; CFI=0,913; RMSEA=0,042 (0,035-0,049)**		
3. Равенство ост. средн.		X2(304)=979,17; CFI=0,664; RMSEA=0,079 (0,074-0,085)		
	Неэквивалентные	15[X2=117] 6[X2=56] 10[X2=15] 5[X2=10]	3[X2=55]	16[X2=102] 5[X2=59] 8[X2=21] 14[X2=22] 2[X2=9] 9[X2=7]
	Частичная эквивал.	**X2(293)=473,42; CFI=0,910; RMSEA=0,042 (0,035-0,048)**		

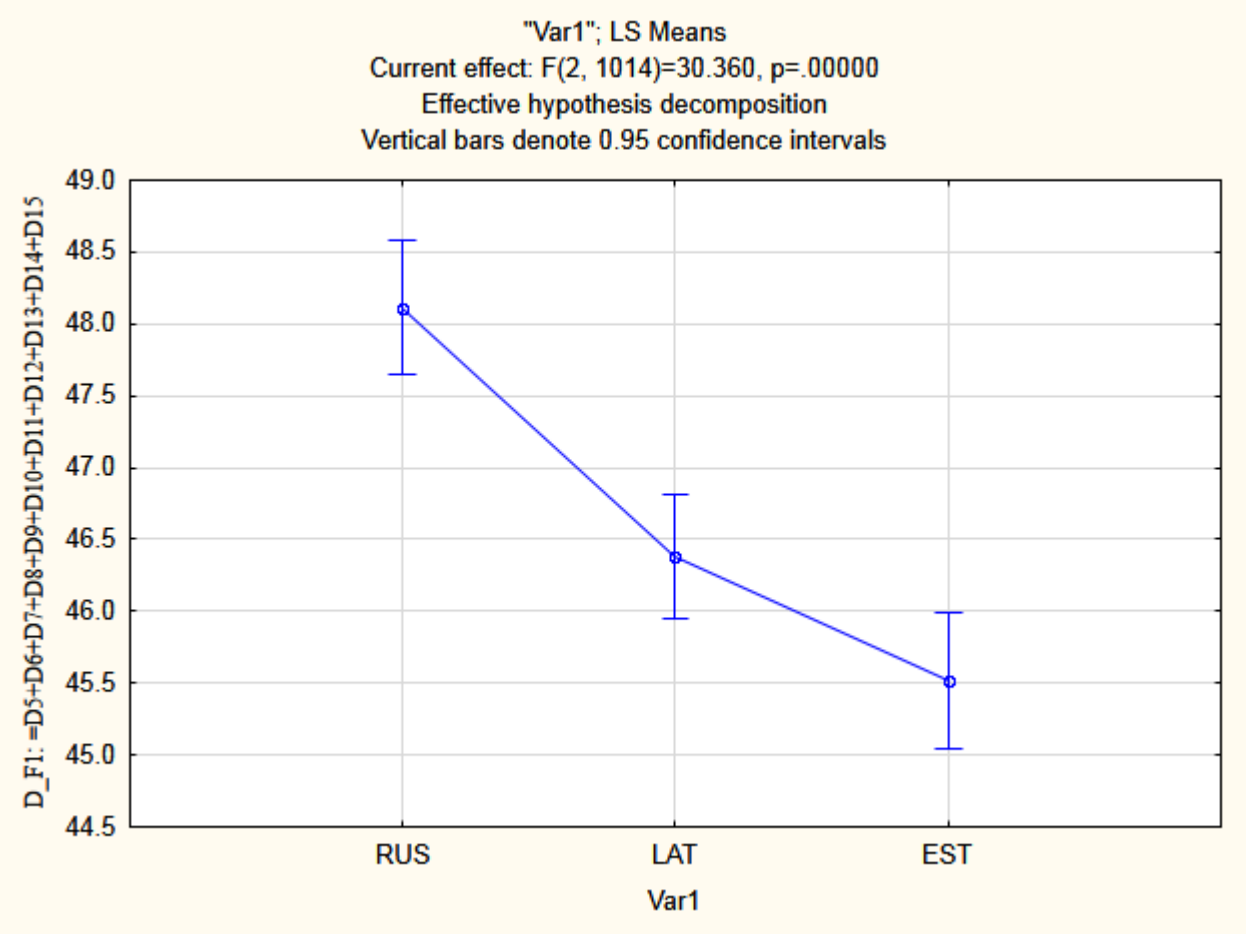
Т.к. априори нельзя утверждать эквивалентность индикаторов, для идентификации модели на этапе 1 дисперсии и средние латентных факторов были приравнены к 1 и 0, соответственно, а все нагрузки и остаточные средние были свободными параметрами. На каждом этапе в модели снимались ограничения до тех пор, пока она не переставала значимо отличаться от итоговой модели предыдущего этапа (выделены).

Показатели латентных факторов

	Россия	Латвия	Эстония
F1 – Конструктивизм среднее	0	-0,08	0,01
F2 – Традиционализм среднее	0	0,03	-0,10
F1 – Конструктивизм дисперсия	1	1,16	1,12
F2 – Традиционализм дисперсия	1	1,21	0,63
Корреляция F1-F2 (станд.)	-0,04	**-0,25 (p < 0,001)**	-0,03

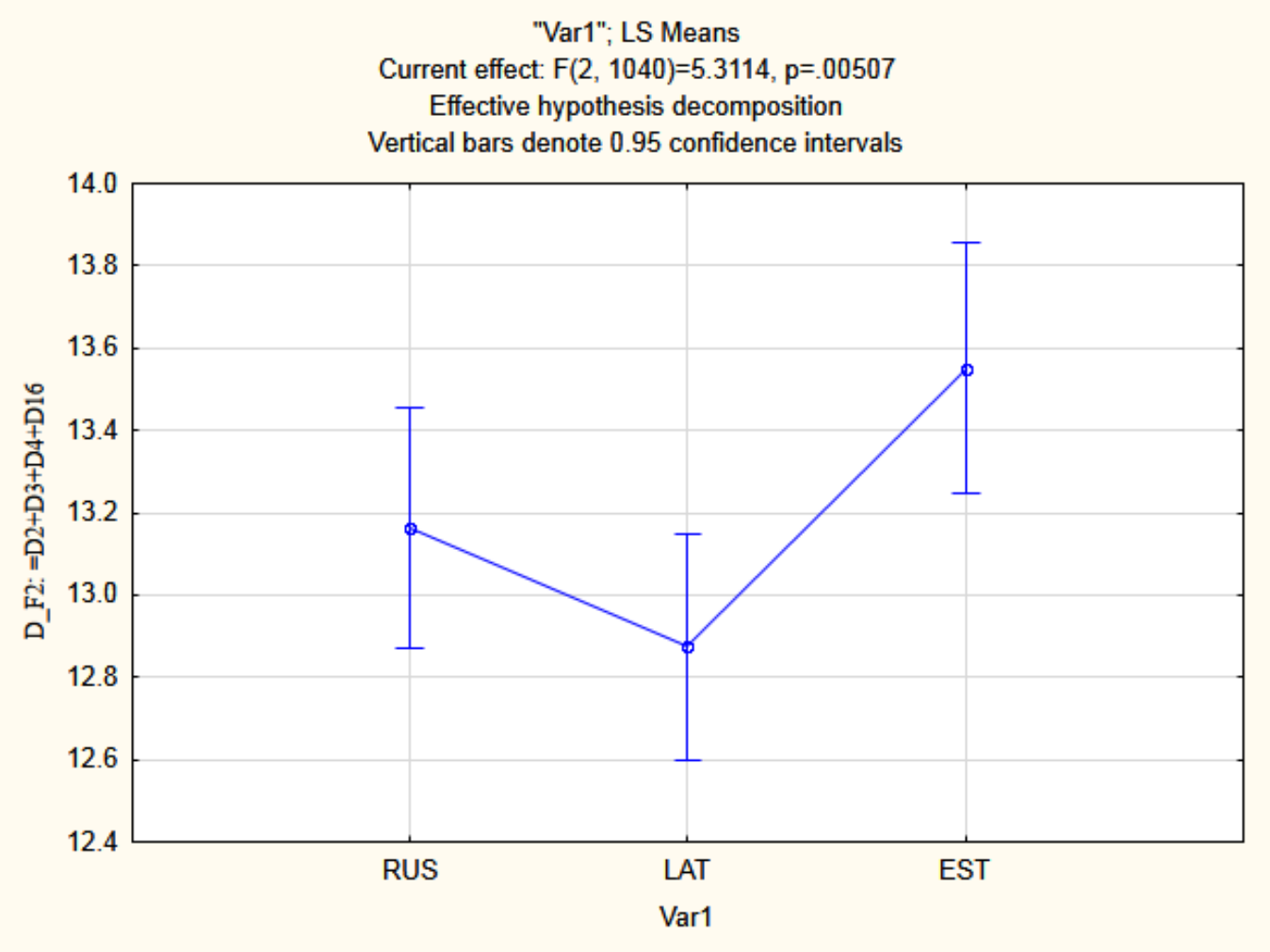
- Учителя в России не отличаются от учителей в Латвии и Эстонии по среднему уровню традиционализма и конструктивизма.
- Но в Латвии эти установки свойственны скорее разным учителям.
- Если не снимать ограничения на равенство остатков переменных 2 и 9, то в Эстонии уровень традиционализма оказывается значимо ниже (станд.: -0,26, p<0,05). Но с учётом того, что в этой стране в отдельности показатели соответствия были наихудшими, это, вероятно, артефакт.
- **Вывод: учителя в России, Латвии и Эстонии не отличаются по установкам.**

А что будет, если просто сравнивать суммарные баллы?



F1: Конструктивизм

А что будет, если просто сравнивать суммарные баллы?



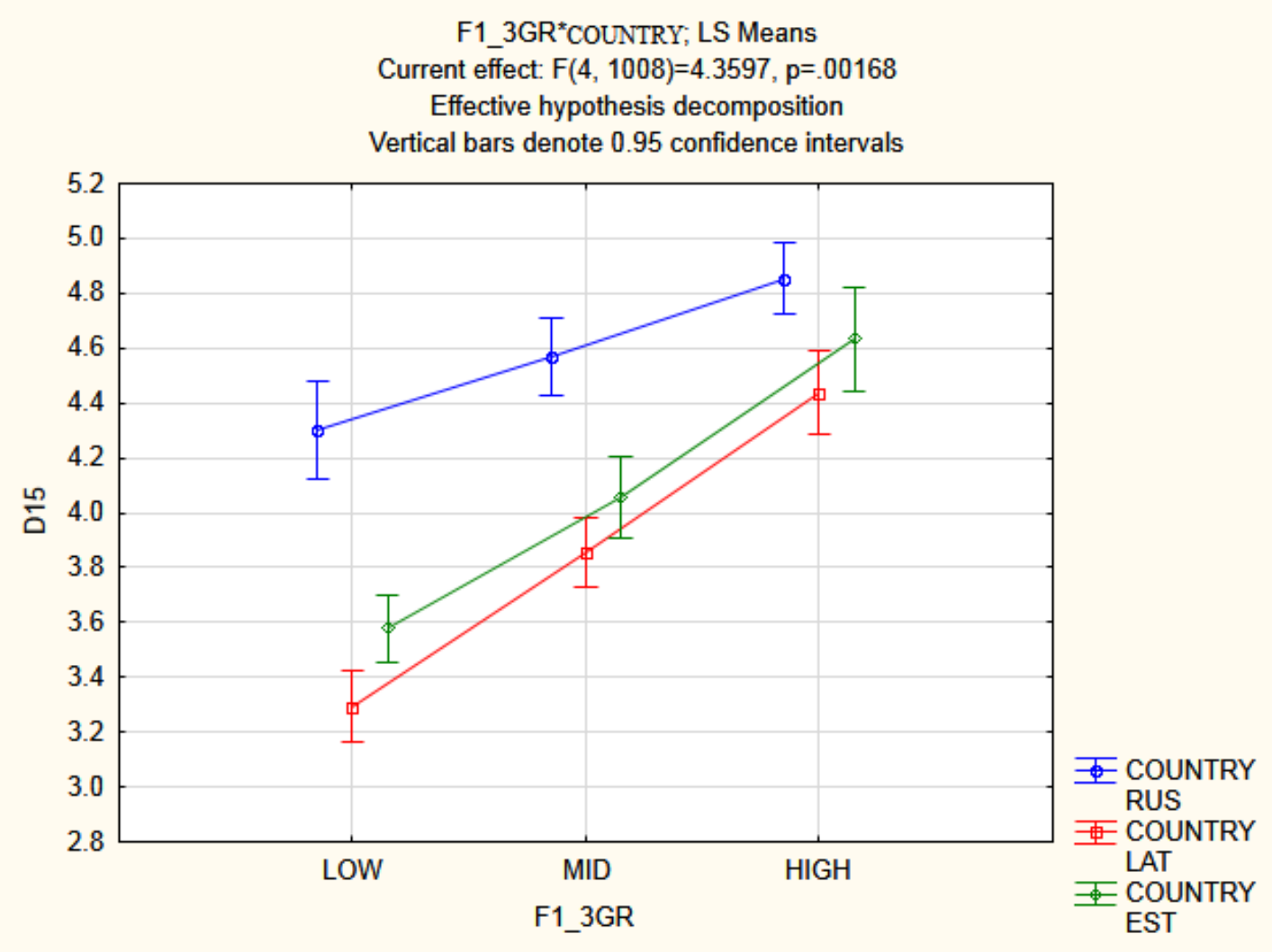
F2: Традиционализм

А что будет, если просто сравнивать суммарные баллы?

- В этом случае мы сделаем выводы (post-hoc Bonferroni test) о том, что:
  - учителя в Эстонии и Латвии – меньшие конструктивисты, чем в России ($p<0,001$);
  - учителя в Эстонии – большие традиционалисты, чем в России ($p<0,01$).
- **Почему?**

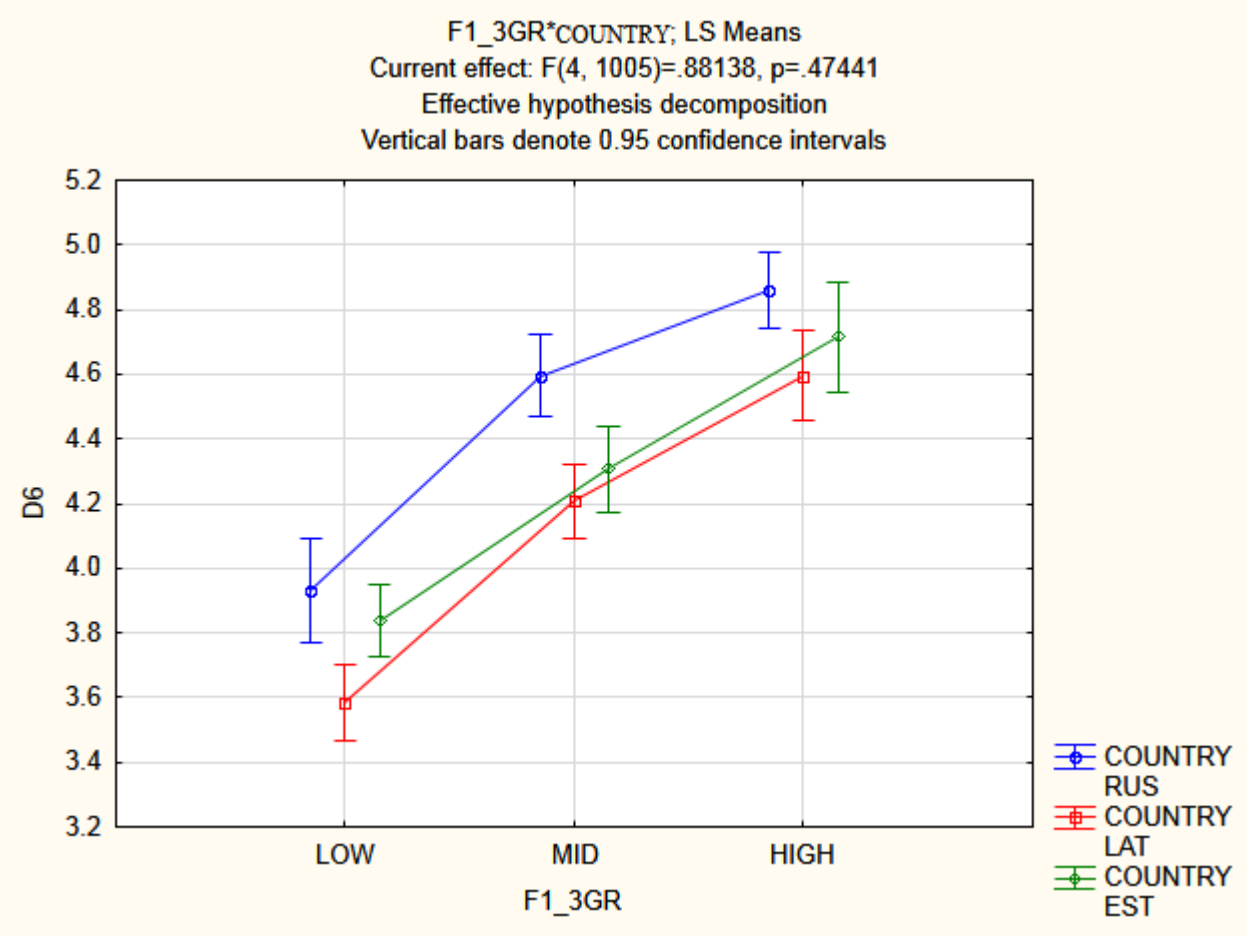
Найдём источник bias

- Посмотрим, как учителя отвечают на неэквивалентные пункты.
- Для этого разобьём учителей на 3 равные по численности (низкую, среднюю и высокую) группы по обеим шкалам, взяв квантили с общей выборки.
- Будем смотреть на пункты, неэквивалентные в России, т.к. с английской версией мы сможем сопоставить только русский вариант перевода.



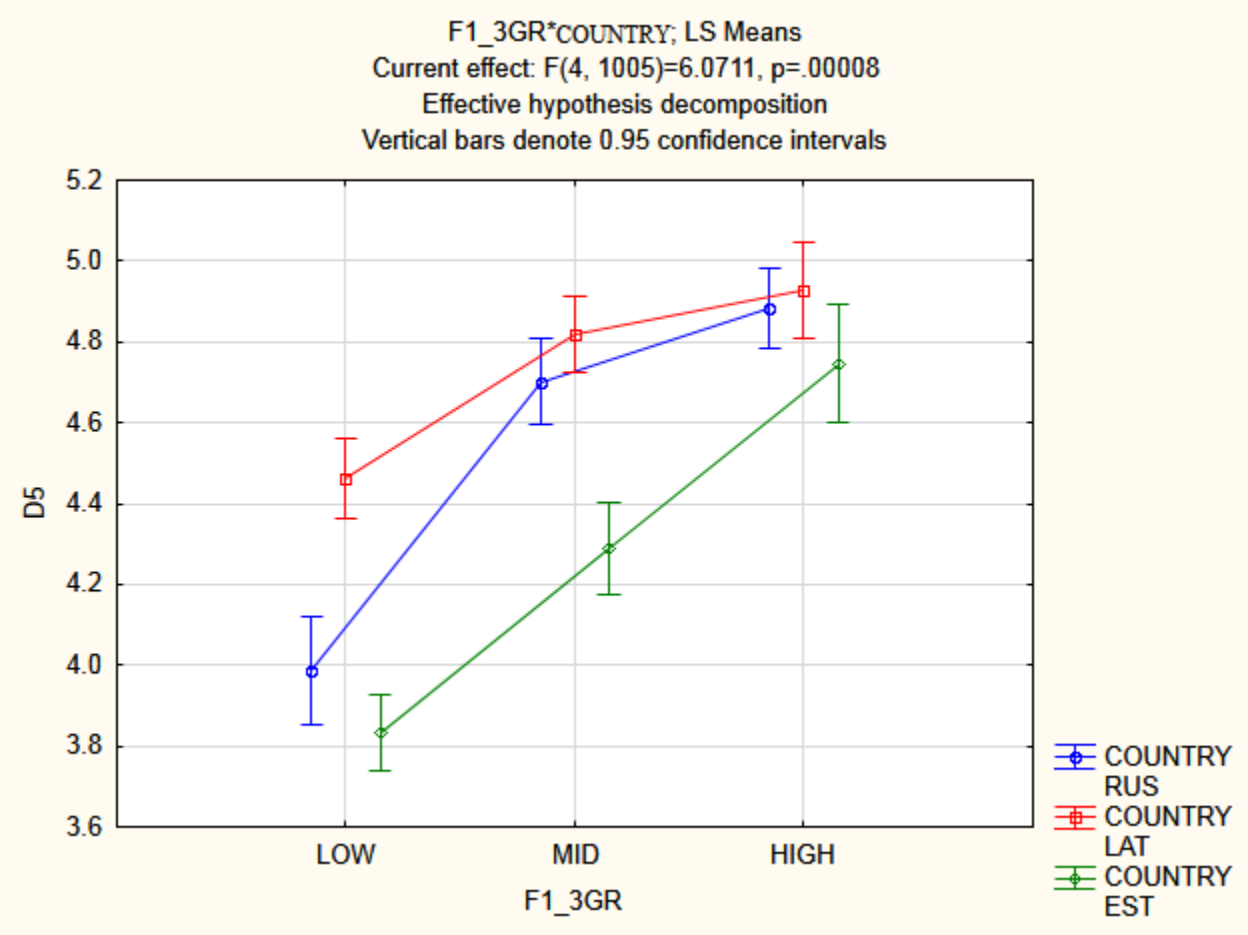
15. Оцениваться должны и практические задачи, проекты, исследования	15. Assessment should include practical problems, projects and investigations.

→ «Оценка должна складываться с учётом результатов работы ученика над…»



6. Ученики учатся лучше всего тогда, когда самостоятельно находят решения заданий	**6. Students learn best by finding solutions to problems on their own.**

→ «Ученики лучше всего учатся путём самостоятельного поиска решений задач»

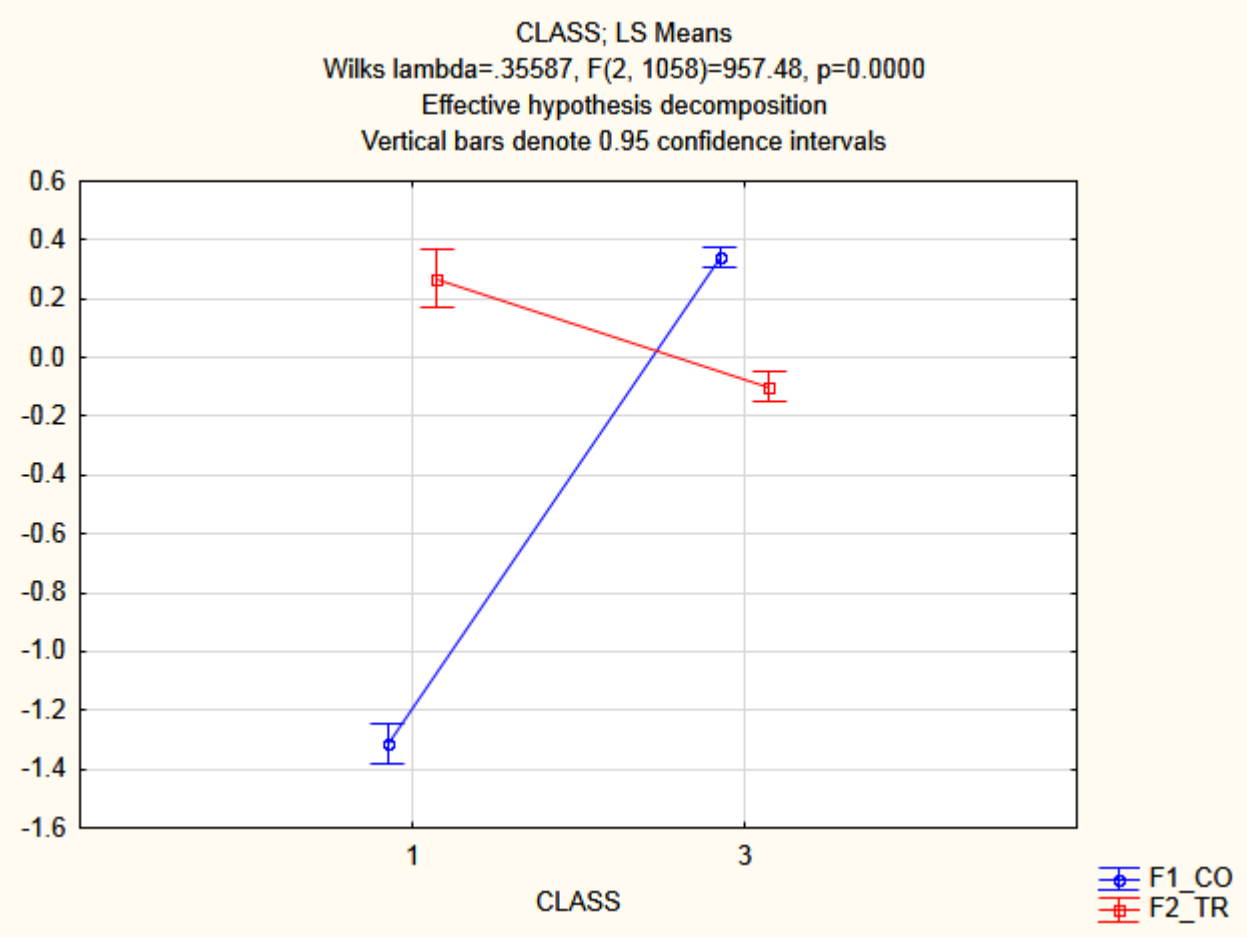


5. Роль учителя – способствовать исследовательской деятельности учеников	**5. My role as a teacher is to facilitate students' own inquiry.**

→ «Моя задача как учителя – способствовать самостоятельной…»

Выход за пределы КФА

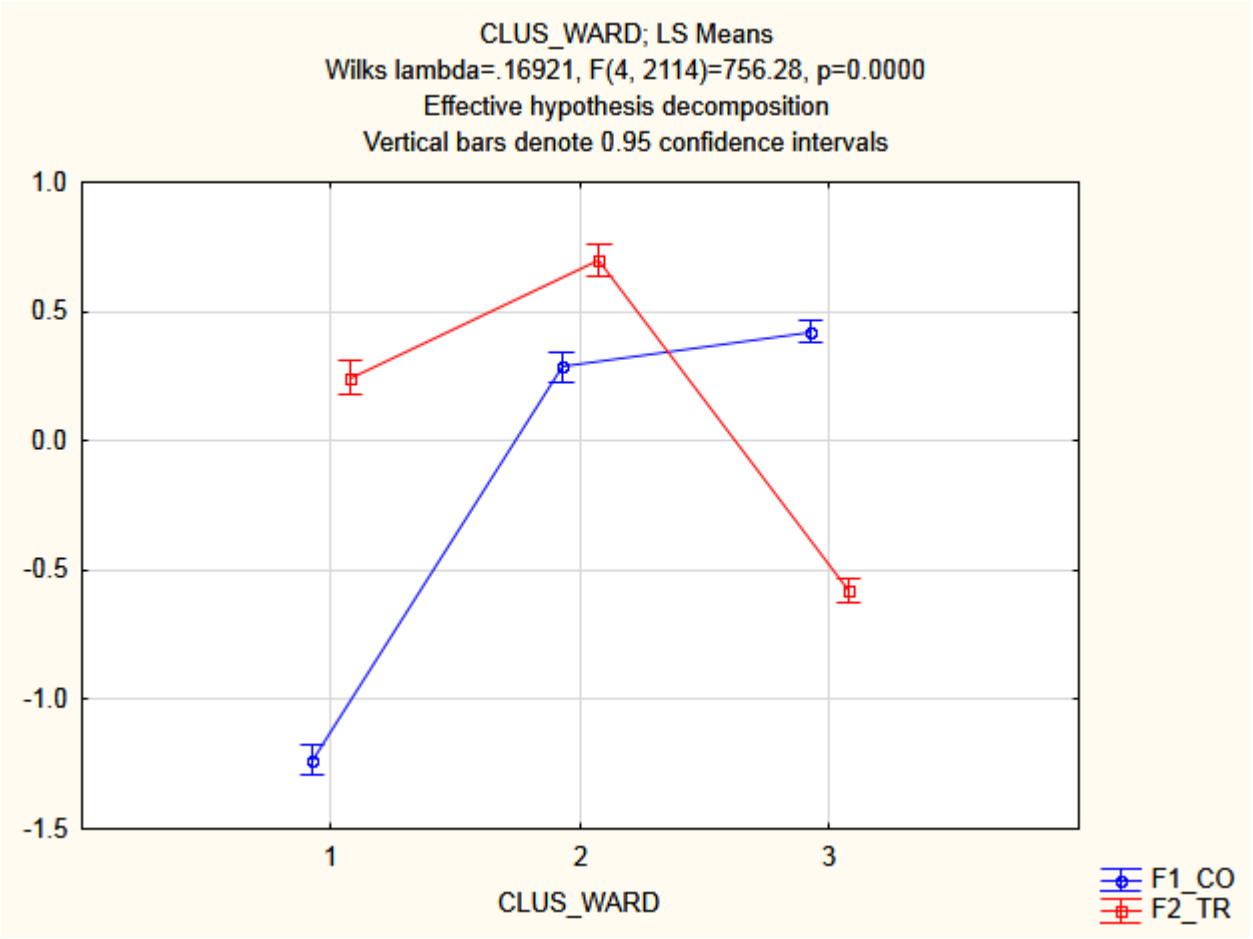
- Теперь мы можем работать с инвариантными оценками респондентов по латентным факторам как с единым массивом данных (экспорт factor scores / plausible values).
- Стыковка с другими методами:
  - например, методы классификации: анализ латентных профилей или кластерный анализ.



Типология учителей на общей выборке

Анализ латентных профилей по факторным оценкам (plausible values) для 2 факторов → 2 класса:

Традиционалисты (1, N=224)
Умеренные конструктивисты (3, N=837)
(но по странам нет различий)



Метод Уорда, квадр. Евкл. метрика

X2(4)=8,07, p=0,089	1. Ярые традиционалисты	2. Эклектики	3. Умеренные конструктивисты
Россия	20,82%	33,14%	46,04%
Латвия	26,29%	27,32%	46,39%
Эстония	23,80%	24,70%	51,51%

Рекомендации по разработке анкет для кросс-культурных исследований

- Априорное определение измеряемых конструктов.
- Планирование количества и соотношения пунктов с разными содержаниями для идентификации каждого фактора (не менее 3 утверждений на фактор).
- Ясная формулировка утверждений, отсутствие сложной лексики и двойных мыслей в утверждениях.
- Двойной (прямой-обратный) перевод с ревизией комитетом экспертов-билингвов.
- Крайне желательно проводить пилотаж анкеты, чтобы убедиться, что ожидаемая факторная структура воспроизводится.

Некоторые ограничения КФА

- Достаточно жёсткая многомерная модель → высокая точность результатов, но трудно обеспечить соответствие модели реальным данным:
  - можно использовать BSEM: КФА на основе Байесовских моделей (distribution-free), approximate measurement invariance.
- Идеальное соответствие показывают лишь короткие инструменты с высокой однородностью пунктов → сужение репрезентации конструктов, ниже валидность.
- Взаимозависимость параметров и их нестабильность на малых выборках → желательны большие выборки, больше культурных групп.
- Трудоёмкость анализа, растущая с колич-вом параметров:
  - эксплораторный подход (alignment approach) к выделению инвариантных параметров: Muthén & Asparouhov, 2013, in press.

Спасибо!

evgeny.n.osin@gmail.com, e_kardanova@mail.ru