НОВ БЪЛГАРСКИ УНИВЕРСИТЕТ
ДЕПАРТАМЕНТ КОГНИТИВНА НАУКА И ПСИХОЛОГИЯ

Критични условия за наблюдаване на ефекта на обърнатата честота

АВТОРЕФЕРАТ
на дисертационен труд за присъждане на образователна и научна степен „доктор“ професионално направление 3.2. Психология научна специалност „Обща психология“

Йолина Петрова

Научен ръководител: доц. д-р Пенка Христова

СОФИЯ, БЪЛГАРИЯ • 2023

Дисертационният труд е обсъден и насочен за защита пред научно жури на заседание на департамент „Когнитивна наука и психология”, Нов български университет, проведено на 19.05.2023 г.

Тезата е предадена на английски език и има следните основни характеристики:
Обем на основния текст: 97 страници
Литература: 120 източника
Графики: 28
Таблици: 20
Приложения: 8

Table of Contents

Въведение

Стремежът да разберем какво стои в основата на способностите ни да организираме знанията си (*понятия*) и да ги използваме (*категоризиране*) води до огромен емпиричен и теоретичен интерес през последните 50 години. В резултат на което „класическият възглед“, който описва понятийната организация и категоризация като управлявани от правила и датира от две хилядолетия, най-накрая е поставен под съмнение в полза на по-динамични възгледи – като подходите, базирани на сходство и на знания (Smith & Medin, 1981). Този научен прогрес е признат за една от „успешните истории“ в когнитивната наука (Gardner, 1985; Gurova, 2013).

Един от приносите на тази „успешна история“ е осъзнаването, че за хората е изключително лесно да научат кои неща се асоциират с опция *А* и кои с опция *Б*. Именно това умение стои в основата на една от класическите когнитивни задачи, наречена задача за класификация. Чрез отгатването на категориите на поредица от примери, задачата изисква от участниците да усвоят даден набор от категории. Тъй като всеки отговор е последван от коригираща обратна връзка, постепенно се навлюдава подобрение в категоризирането на нови примери (Goldstone et al., 2018). Същата тази традиция за учене чрез класификация разкрива един озадачаващ феномен, наречен “*ефект на обърнатата базова честота*” (*ИБРЕ)* (Medin & Edelson, 1988).

1. Ефектът на обърнатата базова честота (*ИБРЕ*) в своята същност

Класическата парадигма, с която се наблюдава *ИБРЕ,* е следната. Хората се инструктират, че трябва да научат две категории (или *болести*) – *А* и *Б.* По-конкретно, това което трябва да научат, е кои характеристики (или *симптоми*) се асоциират с коя категория. Първоначално участниците не са наясно с правилните отговори, поради което започват с отгатване. Тъй като всеки отговор е последван от обратна връзка дали е бил правилен или не, след няколко отгатвания хората се научават да реагират с категория *А*, когато видят „*болки в ушите*, *кожен обрив*“; и да избират категория *Б*, когато видят „*болки в ушите*, *болки в гърба*“ (Kruschke, 1996). Две особености остават имплицитни за участниците. Първо, категориите не се появяват с еднаква честота. Една от тях се появява три пъти по-често от другата (т.е. има често срещана категория и рядко срещана категория). Второ, всяка от категориите се дефинира чрез две характеристики – (1) обща характеристика, срещаща се и в двете

категории (в примера по-горе това е „*болки в ушите“*); и (2) уникална характеристика, която предсказва само една от категориите, т.е. „*кожен обрив*“ за категория *А* и „*болка в гърба*“ за категория *Б*.

Непосредствено след фазата на учене следва и фаза на тестване. Тестовата фаза започва с инструкции, че ще последват нови примери, но задачата остава същата – участниците трябва да продължат да класифицират това, което виждат, в една от категориите, които са научили – вече без обратна връзка. Обикновено предпочитанията за генерализиране на наученото се тестват по няколко начина. Когато на участниците се представи само едно уникално свойство (като “*кожен обрив*” или “*болка в гърба*”), предпочитанията им са очаквани – избират категория *А* в първия случай и *Б* във втория. В това няма нищо изненадващо, тъй като този тип тестови примери не съдържат никаква информация, която би предположила принадлежност към друга категория. Тестовите примери, състоящи се само от общата характеристика (като „*болки в ушите*“) обикновено се класифицират като принадлежащи към по-честата категория, което е в съответствие с базовата честота на категориите. Трите характеристики, представени заедно, (т.е. „*болки в ушите*, *кожен обрив*, *болки в гърба*“) също се класифицират като примери на по-честата категория, макар и не толкова често, колкото предишния критичен тест. От решаващо значение е, когато двете уникални характеристики се показват едновременно (напр. „*кожен обрив*, *болки в гърба*“). В този случай предпочитанието за класификация противоречи на информацията за базовата честота на категориите и примерът се класифицира като принадлежащ към по-рядката категория (Kruschke, 1996). С други думи ефектът е свързан със странно, но консистентно поведение от страна на участниците – в една и съща тестова фаза, в зависимост от тестовия пример, участниците избират да се съобразят с честотната информация на категориите (т. нар. *Общ* тест), почти да я игнорират (т. нар. *Всички заедно* тест) или да се противопоставят на нея (т. нар. *Комбиниран* тест). Точно това обръщане на предпочитанията прави ефекта толкова озадачаващ.

2. Откъде идва важността на ефекта на обърнатата базова честота?

2.1. Практическа значимост на ИБРЕ

Едно от важните практически проявления на феномени като *ИБРЕ* се наблюдава в медицината. Игнорирането на честотната информация, свързана с дадени заболявания, от страна на медицинските специалисти води до сериозно надценяване на вероятността за

диагностицирането в определени заболявания (Casscells et al., 1978) и подценяването на други (Bergus et al., 1995). В името на погрешното игнориране на тази информация, западната медицина превръща поговорки като "*Когато чуеш копита зад себе си, не очаквай да видиш зебра*" в мантра, целейки да напомня на медиците винаги първо да изследват най-вероятните клинични състояния и едва след това да обмислят по-екзотичните такива.

2.2. Преодоляване на пропастта между вземането на решения и категоризацията

Отвъд медицинската област, литературата за категоризиране и вземане на решения показва, че дали хората разчитат на честотна информация или не, зависи от начина, по който са усвоили тази информация (Koehler, 1996; Barbey & Sloman, 2007). Когато честотната информация е предоставена под формата на експлицитно резюме, хората по-скоро я игнорират. И обратното, ако тази информация се придобива имплицитно чрез примери, вероятността тя да бъде използвана се увеличава (Gigerenzer et al., 1988). Идеята тук е, че с нарастването на броя на отделните следи, които всеки пример оставя в паметта, се увеличава и вероятността информацията, свързана с тях, да бъде извлечена. Конкретно *ИБРЕ* има забележително място в литературата, тъй като е едновременно: 1) проява на ефект на ученето чрез примери; и 2) предпочитание, противоречащо на базовата честотна информация. Именно поради това обяснението на *ИБРЕ* може да бъде информативно както за това какво характеризира процесите на вземане на решения, така и процесите, свързани с правене на извод в контекста на честотна информация.

2.3. ИБРЕ като предизвикателство за класическите модели за категоризация

От друга страна, ефектът не се предвижда от нито една от традиционните нормативни теории и теории за ученето. Теоремата на Бейс, например, не предлага нормативен принцип, който недвусмислено определя кой е рационално правилният отговор за *Комбинираните* тестови примери (Medin & Edelson, 1988). Подобно, Теорията за контекста, базирана на примери, на Medin и Schaffer (1978), очаква предпочитание към по-честата категория за всички двусмислени примери (включително и за *Комбинираните*, за които хората демонстрират предпочитание към по-рядката категория). За разлика от това, повечето модели за класификация, базирани на прототипи, нямат специфични очаквания, свързани с честотната информация на категориите, тъй като те третират класификацията като процес на

изчисляване на някаква форма на сходство между тестовия пример и прототипите на категории, към които примерът потенциално принадлежи.

Като цяло *ИБРЕ* изглежда важно изследователско приключение, тъй като има отражение в реалния живот. В случай че тези поведения наистина са ирационални, разбирането на ефекта може да доведе до предотвратяване на евентуални пропуски в употребата на честотна информация. В допълнение, тъй като ефектът е на кръстопътя между вземането на решения и категоризацията, той може да бъде информативен и за двете области и да се използва като разграничаващ инструмент между алтернативните когнитивни модели.

3. Теоретични обяснения на *ИБРЕ*

Към момента има две алтернативни обяснения на *ИБРЕ*. Едното от тях обяснява ефекта чрез асоциативни процеси, които се случват по време на ученето (напр., Kruschke 1996, 2009). Другото обяснение приписва ефекта на процеси на разсъждение от високо ниво, протичащи по време на тестовата фаза (напр., Juslin et al., 2001; Winman et al., 2005).

3.1. Обяснение на ИБРЕ, базирано на асоциации

Доминиращото в момента обяснения на *ИБРЕ* приписва ефекта на отделянето на повече внимание и асоциативна сила на опредлени характеристики на категориите (Kruschke, 1996). Това подчертаване се случва по време на самото им учене. Тъй като една от категориите се появява по-често, тя е и първата, която се усвоява. В началото и двете характеристики на по-често срещаната категория получават еднакво внимание и асоциативна сила. Това не е така при по-рядката категория. Тъй като класификацията въз основа на общата между двете категории характеристика води до грешки, когато става въпрос за по-рядката категория, вниманието се измества към уникалната характеристика от същата категория. (Kruschke, 1996, 2009). Всичко това води до различно разпределение на вниманието по отношение на двете категории. При по-често срещаната категория се приема, че е представена и от двете си дефиниращи характеристики (и те получават относително еднакви внимание и асоциативна сила); докато по-рядката категория се представя най-вече от уникалната си характеристика (и тя получава много по-силна асоциативна тежест от общата за двете категории характеристика).

Накратко, ефектът се отдава на вид *представна асиметрия,* като това обяснение на ефекта се инстанцира и в конекционистки модел, наречен *Extended ADIT Model* (*EXIT*) (Kruschke, 2001b).

3.2. Обяснение на ИБРЕ, базирано на правила

Когато става въпрос за *ИБРЕ*, литературата признава потенциална роля и на разсъжденията от по-висок ред (Kruschke, 2003; Johansen et al., 2007; Winman et al., 2003). Едно такова обяснение на ефекта – наречено *извод чрез елиминация* – идва от Juslin et al. (2001). Те припомнят, че *ИБРЕ* може да се дължи на някаква форма на разсъждение от високо ниво, а не на механизми, свързани с ученето. По същество Juslin et al. (2001) твърдят, че категоризацията е процес на търсене на съвпадения, при което всеки нов пример се проверява по отношение на достатъчно съвпадение с представянията на усвоените категории. Най-често първо се проверява по-честотното правило (категория), тъй като то е по-познато. Ако тестовият пример е достатъчно правдоподобен пример на категорията – в контекста на *ИБРЕ* това означава да има по-малко от една различна характеристика с правилото – то се предполага, че е правилното и се категоризира в него (Juslin et al., 2001). Ако примерът е нов, двусмислен и изглежда неправдоподобен – т.е. има повече от една различна характеристика с правилото – правилото се елиминира в полза на по-рядкото такова. Точно елиминирането на по-честото правило е това, което води до предпочитанието на по-рядката категория, когато тествият стимул е от *Комбиниран* тип. Допускането на този възглед е, че всички категории са представени от целия набор дефиниращи характеристики – т.е. има *представна симетрия*, при която и двете категории са представени и от двете си характеристики (общата и уникалната). Обяснението на *ИБРЕ*, базирано на правила, също е формализирано в модел, наречен *Elimination Model* (*ELMO*) (Juslin et al., 2001).

3.3. Емпирична подкрепа за обясненията на ИБРЕ, базирани на асоциации и на правила

Емпиричната подкрепа за очертаните обяснения е доста двусмислена. От една страна, *ИБРЕ* не се наблюдава без обща характеристика (напр., когато категория *А* се дефинира чрез „*болки в ушите*“ и „*кожен обрив*“, а *Б* се дефинира чрез „*болка в гърба*“ и „*гадене*“) (Johansen et al., 2007; Kruschke, 2001a). Това наблюдение е в съответствие с обяснението, базирано на асоциации, тъй като няма причина за изместване на вниманието от някоя от характеристиките, което се свързва с представната асиметрия (Johansen et al., 2007; Kruschke, 2001a). Междувременно, базираното на правила обяснение на ефекта (формализирано в

ELMO) е безразлично към това дали категориите споделят характеристика или не, тъй като приписва ефекта на по-добро заучаване на правилото, представляващо по-често срещаната категория, а не на структурни разлики между формираните правила.

Нещо повече, в съответствие с асоциативно-базирания подход, установено е, че визуалното внимание, измерено чрез потенциали, свързани със събития (*ERPs*) е по-голямо за тестовия случай, представящ само уникалната характеристика на по-рядката категория (в сравнение с тестовия случай, представящ само уникалната характеристика на по-честата категория) (Wills et al., 2014). В същия ред на мисли са и скорошни *fMRI* изследвания. Изглежда, че специфични области в мозъка получават значително повече активация по време на представянето на уникалната за рядката категория характеристика (Inkster et al., 2022). Важно в случая е, че по-силно активиращите се региони се свързват с грешки при очакванията по време на ученето (напр., Fouragnan et al., 2018, както е цитирано в Inkster et al., 2022).

От друга гледна точка, в подкрепа на базираното на правила обяснение е друго скорошно *fMRI* изследване (O'Bryan et al., 2017). O'Bryan и колеги (2017) демонстрират, че по време на *Комбинираните* тестови случаи участниците обръщат повече внимание на уникалната за по-честата категория характеристика, включително и когато избират да категоризират в по-рядката категория. Това наблюдение съвпада с базираното на правила обяснение на *ИБРЕ*, тъй като предполага, че категоризирането в по-редките категории при двусмислените тестови случаи всъщност е свързано с по-голямо внимание към уникалната характеристика на по-честата категория (т.е. по-честото правило се тества първо и се елиминира).

Очевидно е, че емпиричните данни не дават своята неоспорима подкрепа за нито едно от очертаните формализирани обяснения – нито асоциативно базираното обяснение, обясняващо ефекта като ефект на ученето, нито основаното на правила такова, което приписва ефекта на процеси на разсъждение по време на тестовата фаза.

4. Обосновка на текущата работа

Тезата си поставя три главни цели: 1) да изследва систематично (при различни условия) ролята на ученето за възникването на *ИБРЕ*; 2) да изследва алтернативни обяснения на ефекта; и 3) да тества доминиращото обяснение на *ИБРЕ* (обяснението, базирано на асоциации) в условия, при които придобитите категории са представени симетрично. Следвайки тези цели, планирани бяха шест експеримента и една симулация.

Целта на първия експеримент (Експеримент 1: *ИБРЕ* с Учене чрез класификация) бе да установи големината на ефекта с класическата за него класификационна задача, така че да може да се използва като норма за сравнение с по-нататъшните експерименти. Очакването за този експеримент беше да се наблюдава трансфера на знания и предпочитания, свързани с *ИБРЕ* след учене на категории с разлики в честотата 3:1 и наличието на обща характеристика. Стимулите, използвани в този експеримент (прости визуални стимули, конструирани специално за проекта), са използвани и в експериментите, представени по-долу. Централният въпрос на втория експеримент се отнася до това дали ключът към наблюдаването на *ИБРЕ* е наистина представната асиметрия (придобита по време на ученето на категориите) (Kruschke, 1996, 2009). Предишни изследвания в областта на категоризацията подчертава разликите в представянето на категориите, формирани чрез учене чрез класификация и други видове учене (Chin-Parker & Ross, 2004; Sweller & Hayes, 2010; Yamauchi & Markman, 1998) и по-конкретно чрез ученето чрез извод. Поради това, във втория експеримент (Експеримент 2: *ИБРЕ* с Учене чрез извод) беше използвана задача за учене чрез извод с цел придобиването на симетрично представяне и на двете категории и в същото време се използва стандартната тестова процедура, при която се наблюдава *ИБРЕ*. Мотивите бяха, че ако ефектът все още се наблюдава, тогава ще има силен довод да се преразгледа твърдението, че асиметрично представяне е необходимо условие за наблюдението на *ИБРЕ* (напр., Kruschke, 1996).

Два други експеримента – Експеримент 3: *ИБРЕ* с Мотивация преди ученето и Експеримент 4: *ИБРЕ* с Мотивацията преди тестването – изследваха ефекта в условията на допълнителна мотивация. Според литературата има редица начини, по които мотивацията може да повлияе на когнитивните процеси – напр., 1) достъпност на свързаните с целта понятия, знания и отделни елементи; и 2) общо представяне и учене. Допускането беше, че ако големината на *ИБРЕ* в първия експеримент се различава от големината на ефекта в тези два експеримента, взимайки предвид посоката на ефекта, можем да заключим дали ефектът е модулиран от процеси, свързани с ученето или е по-вероятно да са свързани с фазата на тестване.

Петият експеримент (Експеримент 5: *ИБРЕ* без Учене) елиминира изцяло процеса на постепенно асоциативно учене и *ИБРЕ* беше тестван в задача за вземане на решения. За тази цел фазата на учене беше напълно елиминирана за сметка на въвеждане на обстановка, при която съответната честотна информация за каегориите се предоставя в рамките на един опит за категоризиране (т.е. във всеи опит бяха едновременно презентирани общо 4 примера от целевите категории заедно със стимула, който изисква категоризация). Обосновката зад тази

манипулация беше, че ако ефектът все още се наблюдава, тогава ще има сериозен мотив да се преразгледа твърдението, че *ИБРЕ* е ефект на ученето (Kruschke, 1996).

Допускането за задължително ефективно учене на целевите категории беше изследвано в последния експеримент (Експеримент 6: *ИБРЕ* с Контролно условие). В допълнение данните от последния експеримент бяха изследвани по-детайлно, предлагайки някои нови интересни резултати (напр., фактът, че участниците, които не покриват критерия за учене, все още демонстрират предпочитания, подобни на *ИБРЕ*). В експеримента беше добавено и контролно условие, тестващо твърдението, че честотната разлика по време на ученето чрез примери, е необходимо условие за наблюдаването на *ИБРЕ* (Kruschke, 1996). Участниците трябваше да научат две двойки категории – едната двойка следваше класическото съотношение 3:1 между категориите, докато в другата двойка и двете категории се появяваха еднакво често. Аргументът тук беше, че ако *ИБРЕ* се наблюдава в двойката с честотни разлики, но не и в контролната двойка, действително ще имаме причина да отдадем ефекта на наличието на честотни разлики между двете категории, споделящи припокриваща се характеристика, а не на други замърсяващи фактори.

В допълнение, обясненията, базирани на асоциации и базирани на правила, бяха изследвани и чрез експлицитно косвено измерване на структурата на придобитите категории. В последните фази на Експеримент 1: *ИБРЕ* с Учене чрез класификация и Експеримент 2: *ИБРЕ* с Учене чрез извод, участниците бяха помолени да опишат какво определя всяка от категориите, които са научили. Събраните вербални протоколи позволиха да се изследва дали има представна асиметрия (както се предполага от подхода, базиран на асоциации) и приоритизиране на по-честата категория (както се предлага от подхода, базиран на правила).

Накрая е представена компютърна симулация на *ИБРЕ*. Целта на симулацията беше да се проучи дали ефектът може да бъде наблюдаван с чисто асоциативна и базирана на вероятности архитектура като авторегресивния езиков модел от тип трансформатор без допълнително учене (по-конкретно, *GPT-3*, Brown et al., 2020). Обосновката беше, че ако ефектът се наблюдава с такъв модел, налагането на необходимост от каквито и да било допълнителни представни асиметрии, усвоени по време на ученето, би било неразумно, тъй като симулацията не включва допълнително учене, което да промени предварителните представяния на модела.

5. Експеримент 1: *ИБРЕ* с Учене чрез класификация

Обосновка на Експеримент 1

Две цели бяха поставени с този експеримент – 1) да се установи норма за големината на *ИБРЕ* с прости визуални стимули, 3:1 честотна разлика и набор от инструкции, които да послужат като основа за сравнение с последващите експерименти; и 2) да изследва експлицитните знания на участниците за структурата на придобитите категории и всяко потенциално приоритизиране, свързано с тях. Като цяло, първият експеримент следва процедурата на експеримент, докладван от Kruschke (1996, Експеримент 1) – всички участници трябваше да научат четири категории в лабораторна среда (две двойки категории с по две характеристики, като всяка двойка споделя обща характеристика и съдържа по-честа и по-рядка категория). Експериментът се различава от този на Kruschke (1996) в използвания стимулен материал и опита да оцени експлицитното знание на участниците за усвоените категории. За всички експерименти, представени в дипломната работа, са създадени прости и добре контролирани визуални стимули, които са лесни за вербализиране, така че да е малко вероятно придобитите представяния и предпочитанията за генерализиране да са повлияни от предишни знания, перцептивна яркост и т.н. Експериментът също така използва процедура, разработена за изследване на съзнателния статус на експлицитното знание на участниците относно структурата на придобитите категории. Процедурата беше използвана като косвена мярка за приоритизирането на категориите и приоритизирането на самите характеристики.

Участници

Общо 70 участници взеха участие в експеримента в замяна на частичен кредит за курса. Осем от тях бяха изключени от анализа, тъй като в третия и последен блок от фазата на учене постигат по-малко от 70% правилни отговори или за честите категории, или за редките (или и за двете). Крайната извадка се състоеше от 62 участници (средна възраст = 23.9 години, SD = 8.8, 48 жени). Един от тези участници не беше взет предвид за вербалната част на експеримента поради техническа грешка при събирането на устния доклад на този участник.

Материали

Свойствата, използвани са конструирането на стимули, включваха 4 цветни квадрата (*червено, циан, синьо* и *жълто*) и 4 черни фигури (*сърце, кръг, звезда* и *триъгълник*), Фигура

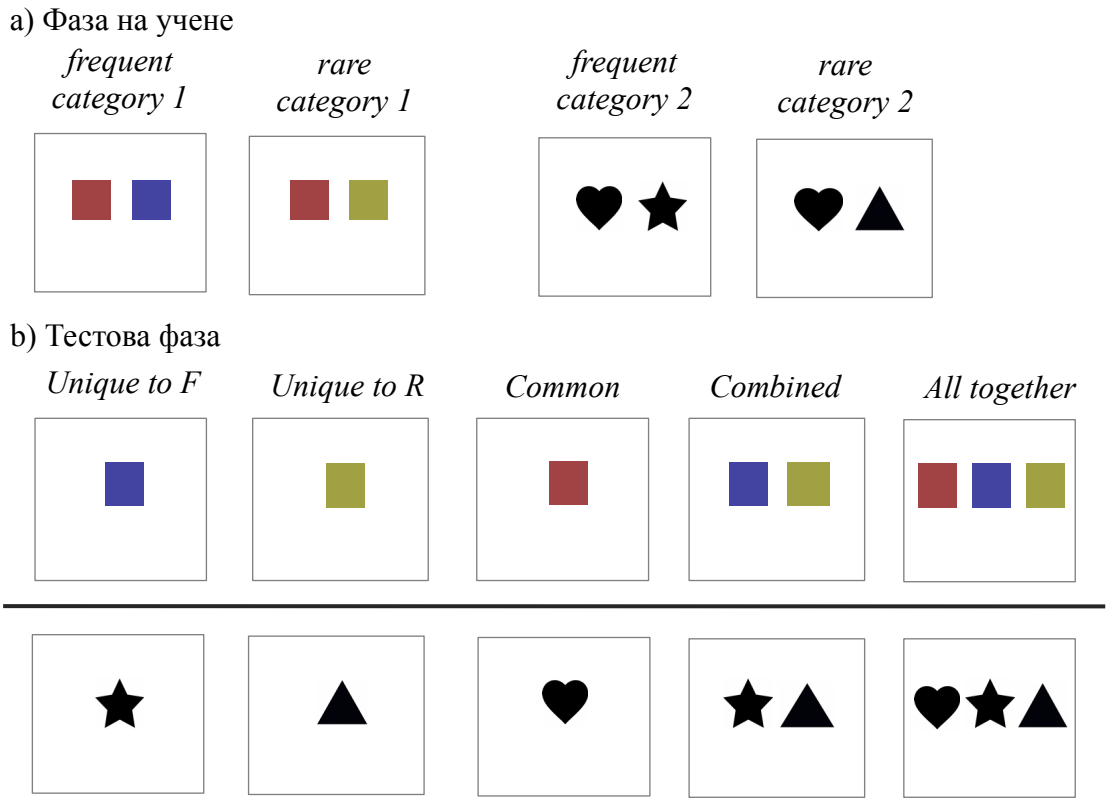
2 за референция. Цветовете бяха избрани от т.нар. Тетрадични цветове, равномерно разпределени по цветното колело, което гарантира, че няма доминация на нито един от тях.

Фигура 2. Всички свойства за изграждане на категориите от Експеримент 1 до Експеримент 6. Отляво надясно – *синьо, червено, синьо, жълто, кръг, сърце, звезда* и *триъгълник*.

За всеки участник бяха проектирани две двойки категории с припокриващи се характеристики (двойките съдържаха една честа и една рядка категория). На случаен принцип за всеки участник една от двойките в категорията беше представена от цветни квадратчета, а другата от двойка с черни фигури. Всяка категория в двойката беше определена от две характеристики (или два цвята, или две фигури) (за пример: Фигура 3, а) – една характеристика, която е уникална за категорията, и втора, която е обща за категориите в съответната двойка. Важно да се отбележи е, че пространствената позиция на характеристиките една спрямо друга не бе от значение.

Процедура

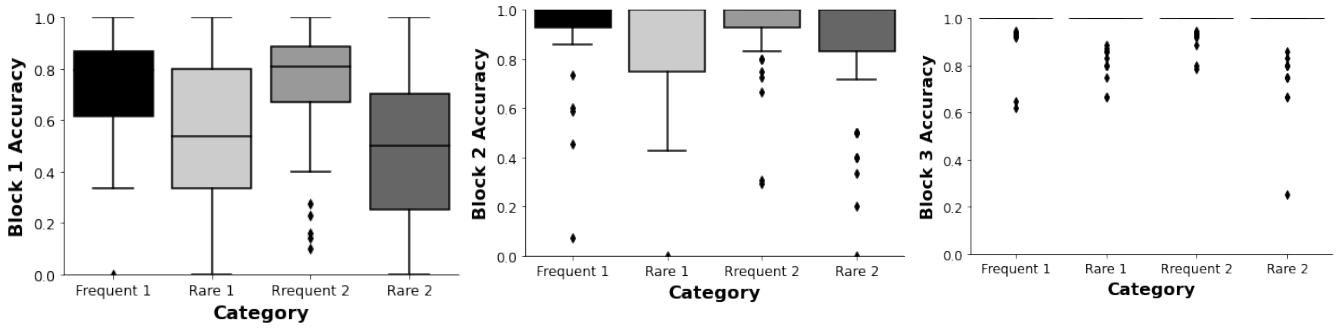
Експериментът се състоеше от три фази: фаза на учене, фаза на тестване и фаза за събиране на вербални протоколи (Фигура 3). Преди фазата на учене всички участници получиха писмени инструкции, информиращи, че ще видят различни изображения и за всяко от тях трябва да отговорят към коя от четири категории принадлежи – "*V*", "*B*", "*N*", или "*M*", натискайки съответния QWERTY клавиш. Групирането на категориите не беше изрично посочено на участниците в нито един момент. Единственото, за което участниците бяха уведомени е, че всеки техен отговор ще бъде последван от коригираща обратна връзка. По време на фазата на учене всички участници бяха презентирани с общо 120 учебни опита, представени в случаен ред. Опитите по време на ученето (примери могат да бъдат видяни на Фигура 3, панел *а*) бяха разделени по следния начин: по-честата категория беше представена 45 пъти, а съответстващата ѝ по-рядка – 15 пъти. С други думи, и в двете двойки категории една от категориите се появява три пъти по-често от другата (т.е., съотношение 3:1). След всеки отговор стимулът изчезваше от екрана, последван от писмена обратна връзка за 1000 ms ("Правилно!" в зелено или "Грешно!" в червено, в зависимост дали отговорът е бил правилен или грешен). Всяко условие започва с фиксационно кръстче, представено за 500 ms и завършва с 1000 ms интервал между условията (ITI).



Фигура 3. Трите фази в Експеримент 1: *ИБРЕ* с Учене чрез класификация. Честотата на категорията и типът на критичния тест са написани над стимулите за яснота на дизайна на експерименталните стимули и процедурата, но не са показват по време на експеримента. Първият ред от критични тестови стимули е за стимули с цветове като характеристики, а вторият ред – с фигури като характеристики.

Преди тестовата фаза участниците бяха информирани, че задачата ще остане същата (класификация в “*V*”, “*B*”, “*N*”, или “*M*” категория), но с нови примери на току що научените категории и без обратна връзка. Инструкциите бяха последвани от 20 тестови опита (по 4 на критичен тест, Фигура 3, панел *b*). Експериментът завърши с молба към участниците да изброят устно какво определя всяка от четирите категории, които са научили в началото на експеримента (за точна формулировка на инструкцията вижте Фигура 3, панел *c*).

Резултати и Дискусия



Фигура 4. Средна точност ст. отклонение по категория в Блок 1, Блок 2 и Блок 3 отляво надясно за Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Учене. Следвайки Kruschke (1996, Експеримент 1), 120-те учебни опита бяха разделени в три блока по 40 опита всеки. Делът на верните отговори в първата част на ученето (първите 40 от 120 опита) беше по-висок за по-честите категории (0.74), в сравнение с по-редките (0.50). Очевидно по-честите категории са придобити по-бързо от редките ($t(61) = 6.68$, $p < .001$, $d = 0.864$ с 95% *CI* [0.17, 0.31]), Фигура 4, *a*), както се твърди от асоциативно базираното обяснение на *ИБРЕ* (Kruschke, 1996). Разликата става по-малка във втората част от ученето ($t(61) = 3.65$, $p < .001$, $d = 0.463$ с 95% *CI* [0.04, 0.12]). До края на третата и последна част от фазата на учене тази разлика на практика изчезва (0.98 за по-честите и 0.96 за по-редките категории), въпреки че остава значима – ($t(61) = 2.67$, $p = .010$, $d = 0.342$ с 95% *CI* [0.01, 0.04]), Фигура 4, *c*). Както може да се види на Фигура 4, по-често срещаните категории са научени много по-рано от редките. И все пак до края на учебната фаза всички категории са добре усвоени.

Тестване. Както се очаква, хората правилно избират по-честата категория, когато им се представя нейна уникална характеристика (в 88% от случаите); аналогично за уникална характеристика на по-рядко срещана категория (избирайки рядката категория в 82% от случаите), реферирай към Таблица 2. По-важното е, че пропорциите на предпочитанията ясно показват, че свързаното с *ИБРЕ* поведение е успешно възпроизведено. Хи-квадрат анализ, изчислен върху честотите на отговор за *Комбинираните* тестови случаи и очаквани стойности от 50:50, показва малки до средни предпочитания към по-рядката категория – $\chi^2(1, N=237) = 7.09$, $p = .008$, $\varphi = .173$ и 95% *CI* [83, 114] и [124, 154] съответно за честите и редките избори. За *Всички заедно* тестови случаи хората демонстрират лек, но незначителен уклон към базовата честота – $\chi^2(1, N=240) = 3.75$, $p = .053$, $\varphi = .125$ с 50:50 очаквани стойности, 95% *CI* [119, 150] и [90, 121] за честите и редките отговори. И накрая, за *Общите*

тестови случаи хората се придържаха силно към базовите честоти на категориите – $\chi^2(1, N=238) = 87.13$, $p < .0001$, $\varphi = .605$ отново с 50:50 очаквани стойности и 95% *CI* [178, 203] и [35, 60] за честите и редките избори.

Таблица 2: Съотношение на предпочитаните категории по тип тест за Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Test Cases	**Choice proportion**			
	Frequent	**Rare**	**Frequent O**	**Rare O**
Unique to F	**0.88**	**0.06**	**0.03**	**0.03**
Unique to R	**0.11**	**0.82**	**0.02**	**0.05**
Common	**0.76**	**0.19**	**0.01**	**0.04**
Combined	**0.39**	**0.55**	**0.01**	**0.05**
All together	**0.54**	**0.42**	**0.02**	**0.02**

Вербални доклади. Устните доклади, отнасящи се до експлицитните знания на участниците за четирите усвоени категории (Фигура 3, *c*)), бяха кодирани по отношение на типа на докладваната дефиниция на дадена категория. След внимателен качествен анализ, всички докладвани дефиниции бяха групирани в 6 типа – „*пълна, започващо с обща характеристика*“, „*пълна, започващо с уникална характеристика*“, „*само често срещана*“, „*само уникална*“, „*неправилна*“ и „*некласифицирана*“. Като цяло, това, което се наблюдава (обобщено в Таблица 3) е, че често срещаните категории са представени от двете им характеристики, без да се дава приоритет на нито една от тях, докато редките категории са представени главно от техните уникални характеристики. Това наблюдение е в съответствие с подхода, базиран на асоциации (Kruschke, 1996), приписващ ефекта на представни асиметрии.

Таблица 3: Устно докладвани дефиниции на категории за тип дефиниция и тип категория в проценти за Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Type of category	**Reported definition**						
	complete, first common	*complete, first unique*	*common only*	*unique only*	*incorrect*	*unclassified*	**total**
frequent	31.97	35.25	7.38	14.75	5.73	4.92	100
rare	18.03	22.95	4.92	33.61	13.93	6.56	100

6. Експеримент 2: *ИБРЕ* с Учене чрез извод (нужна ли е представна асиметрия за наблюдаването на *ИБРЕ*)

Обосновка на Експеримент 2

През последните две десетилетия сме свидетели на нарастващ интерес към разликите в ученето чрез извод и чрез класификация (Chin-Parker & Ross, 2004; Sweller & Hayes, 2010; Yamauchi & Markman, 1998). Една от забележителните разлики между двете задачи е, че ученето чрез класификация дава приоритет на разграничаващите (уникални) пред общите характеристики на категориите, докато учещите чрез извод са принудени да разпределят вниманието си върху характеристиките на категорията по-равномерно (Sweller & Hayes, 2010). Ученето чрез извод изисква предсказване на характеристика (напр., кое е липсващото свойство – "*кожен обрив*" или "*болка в гърба"*), при наличието на конкретна категория (напр., *категория А*) и друго/и свойство (напр., "*болки в ушите*"). С други думи, участникът е информиран, че даден стимул е пример на *категория А* и има свойство *X*; като задачата е да избере кое е липсващото свойство на съответната категория – *Y* или *Z*. Поради естеството на задачата, приоритизирането на която и да е от характеристиките е недостатъчно за оптимална постижение в задачата, тъй като в различните опити липсващата характеристика може да е както уникалната за дадената категория, така и характеристиката, която е споделена междудвете категории.

Ако представната асиметрия е необходимо условие за наблюдаване на *ИБРЕ*, както твърди базирания на асоциации подход (Kruschke, 1996), тогава ефектът не би следвало да се наблюдава при условия на учене, водещо до симетрични представяния – каквото е ученето чрез извод. Ако при такова учене ефектът все още се наблюдава, тогава може да се заключи, че класическата версия на парадигмата на *ИБРЕ* (използваща учене чрез класификация) наследява представната асиметрия, за която Kruschke (1996) предполага, просто като страничен ефект от това учене, но не е критично условие за проявлението на *ИБРЕ*.

Участници

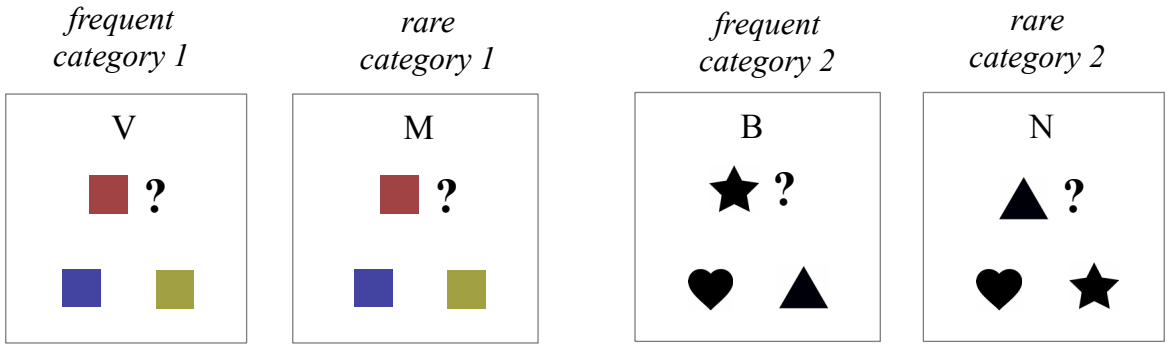
Общо 70 участници взеха участие в експеримента в замяна на частичен кредит за курса. Четиринадесет от тях не бяха включени в анализа, тъй като отбелязаха по-малко от 70% правилни отговори за по-честите, за по-редките или и за двата типа категории. Така крайната извадка се състои от 56 участници (средна възраст = 23.9 години, SD = 6.4, 39 жени).

Материали

Визуалните свойсвто (Фигура 2) и структурите на категориите (Фигура 3, панел а) бяха идентични с тези в Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Процедура

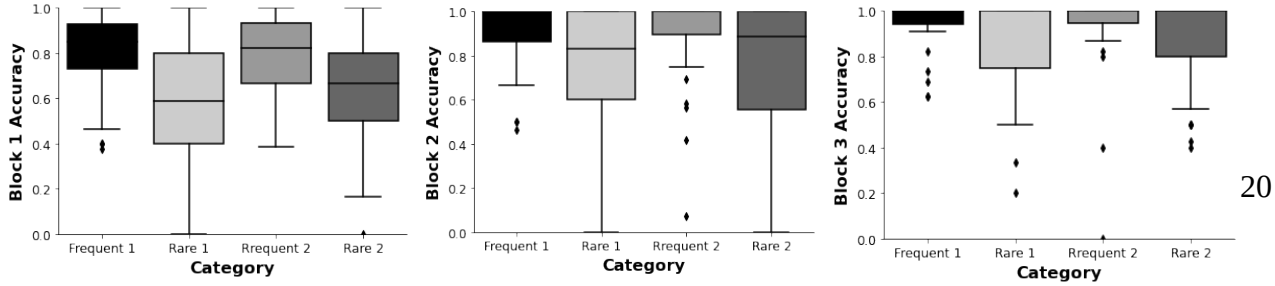
За разлика от предишния експеримент, фазата на учене на този изискваше учене чрез извод. Във всеки опит участниците бяха презентира с една характеристика, която принадлежи към посочена категория, разположена в центъра на екрана в черен контурен квадрат близо до въпросителен знак, който сигнализира за липсваща характеристика (Фигура 7).



Фигура 7. Примерни стимули за опитите във фазата на учене в Експеримент 2: *ИБРЕ* с Учене чрез извод. За пълната структура на категориите вижте Фигура 3, панел а).

Под всеки стимул бяха представени две характеристики, една до друга. Една от характеристиките винаги е липсващата правилна опция, а другата е уникалната характеристика на другата категория от същата двойка. Двете винаги бяха разположени на случаен принцип (една спрямо друга) в долната част на екрана. Важно в случая е, че липсваща характеристика можеше да е както общата за двете категории в двойката, така и уникалната за посочената категория. Това гарантира, че участниците обръщат внимание и на двете характеристики от всяка категория. Задачата на участниците беше да отговорят коя е липсваща характеристика, натискайки съответния бутон : '*Z*' за тази от ляво и '*X*' за опцията, презентирана отдясно. Във всички останали отношения процедурата на експеримента имитираше тази на Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Резултати и Дискусия



Фигура 8. Средна точност и ст. отклонение по категория в Блок 1, Блок 2 и Блок 3 отляво надясно за Експеримент 2: *ИБРЕ* с Учене чрез извод.

Учене. Подобно на Експеримент 1: *ИБРЕ* с Учене чрез класификация, по-честите категории са научени много по-рано от по-редките (Фигура 8). В първия блок делът на правилните отговори за по-честите категории (0.80) е значително по-висок от дела на правилните отговори за по-редките (0.62): $t(55) = 6.83$, $p < .0001$, $d = 0.912$ и 95% *CI* [0.13, 0.24], Фигура 8, *a*). Тази разлика намалява, но остава значителна през третата част от ученето (0.95 и 0.88, съответно за по-честите и за по-редките категории, $t(55) = 4.57$, $p < .001$, $d = 0.616$ и 95% *CI* [0.04, 0.11]), Фигура 8, *c*). И все пак до края на фазата на учене категориите са добре усвоени.

Таблица 4: Съотношение на предпочитаните категории по честота и тип тест за Експеримент 2: *ИБРЕ* с Учене чрез извод.

Test Cases	**Choice proportion**			
	Frequent	**Rare**	**Frequent O**	**Rare O**
Unique to F	**0.78**	**0.12**	**0.07**	**0.03**
Unique to R	**0.17**	**0.74**	**0.05**	**0.04**
Common	**0.61**	**0.27**	**0.06**	**0.06**
Combined	**0.34**	**0.54**	**0.05**	**0.07**
All together	**0.5**	**0.42**	**0.05**	**0.03**

Тестване. Предпочитанията на участниците са в съответствие с *ИБРЕ*. Това е въпреки че задачата за учене чрез извод поощрява симетрични представяния на категориите (Sweller & Hayes, 2010; Yamauchi & Markman, 1998). Таблица 4 показва съотношението на избора за всеки тестов тип. Хи-квадрат анализът, изчислен върху честотите на отговор за *Комбинираните* тестови случаи, показва малки до умерени предпочитания на рядката категория – $\chi^2(1, N=198) = 8.91$, $p = .003$, $\varphi = .212$ с 95% *CI* [64, 92] и [106, 134] за по-честите и по-редките категории съответно. За тестовите случаи от тип *Всички заедно*, хората показват числено предпочитание към честотната категория, макар и статистически незначимо – $\chi^2(1, N=205) = 1.76$, $p = .185$, $\varphi = .093$ и 95% *CI* [98, 126] и [79, 108] съответно за по-честите и по-редките категории. И накрая, за *Общите* тестови случаи хората демонстрират умерено до силно предпочитание към по-честата категория – $\chi^2(1, N=198) = 30.73$, $p < .0001$, $\varphi = .394$ с 95% *CI* [124, 151] и [48, 74] съответно за по-честата и по-рядката опция.

Вербални протоколи. Докладваните дефиниции на категориите бяха групирани в същите 6 типа, както в Експеримент 1: *ИБРЕ* с Учене чрез класификация – „*пълна, започващо с обща характеристика*“, „*пълна, започващо с уникална характеристика*“, „*само често срещана*“,

„*само уникална*“, „*неправилна*“ и „*некласифицирана*“, което позволява изследването на експлицитния знание на участниците относно дефинициите на научените категории.

Таблица 5: Устно докладвани дефиниции на категории за тип дефиниция и тип категория в проценти за Експеримент 2: *ИБРЕ* с Учене чрез извод.

	Reported definition						
Type of category	*complete, first common*	*complete, first unique*	*common only*	*unique only*	*incorrect*	*unclassified*	**total**
frequent	39.29	38.39	2.68	8.93	9.82	0.89	100
rare	39.29	32.14	0.89	15.18	11.61	0.89	100

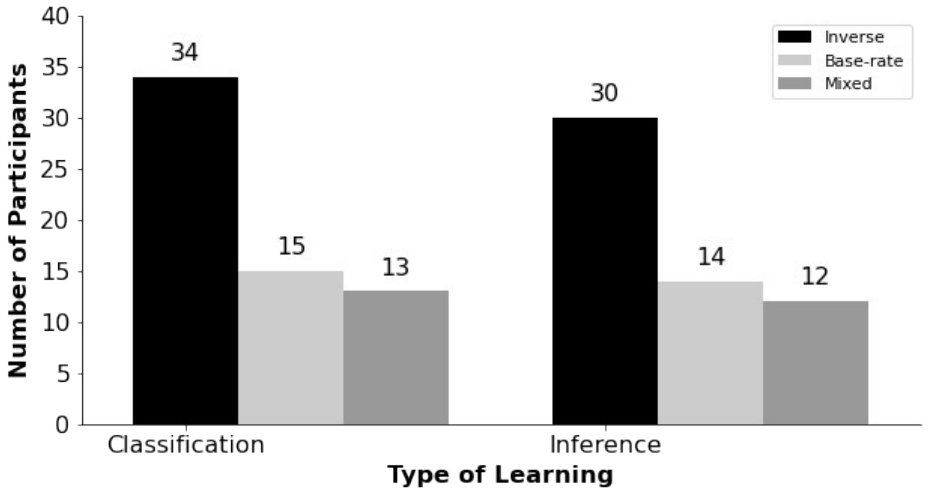
Както се очаква и е показано в Таблица 5, няма значителни разлики между дефинициите на двата типа категории (чести и редки) – $\chi^2(5, N=224) = 3.6$, $p = .608$, $w = .13$. Изглежда, че задачата за учене чрез извод действително води до по-симетрични представяния (както твърдят Sweller & Hayes, 2010; Yamauchi & Markman, 1998 и т.н.). Взети заедно, резултатите остават в контраст както с устно докладваните дефиниции на учещите чрез класификация (от Експеримент 1: *ИБРЕ* с Учене чрез класификация), така и с асоциативно базираното обяснение на *ИБРЕ*, отдаващо ефекта на асиметричното представяне на двата типа категории. По-скоро учещите чрез извод изглежда представят и двете категории по симетричен начин. И въпреки това са подложени на *ефекта на обърнатата базова честота*. Това би могло да означава, че учещите чрез класификаци в парадигмата на *ИБРЕ* действително формират асиметрични представяния с цялостно приоритизиране на уникалната характеристика на по-рядката категория, но това не е критично и необходимо условие, за да се наблюдава *ИБРЕ*. Изглежда много по-вероятно, представната асиметрия да е просто страничен ефект от типа учене, но не и причината за самия ефект.

ИБРЕ при две задачи за учене – сравнение между Експеримент 1 (ИБРЕ с Учене чрез класификация) и Експеримент 2 (ИБРЕ с Учене чрез извод)

Манипулирането на задачата за учене в рамките на процедурата на *ИБРЕ* (чрез използване на учене чрез класификация в първия експеримент и на учене чрез извод във втория) не повлиява големината на ефекта при предпочитанията за класификация на *Комбинираните* тестови случаи, $\chi^2(1, N=435) = 0.17$, $p = .679$, $w = .02$.

Участниците бяха разделени и според начина, по който отговарят на *Комбинираните* тестове (*обърнат, с базова съответствие* и *смесен* тип генерализация). Разделението категорично показва, че ефектът не е единен, т.е. не всички участници са подложени на него (също

отбелязано и от Winman et al., 2005). Това важи за участниците и в двата експеримента (Фигура 11).



Фигура 11. Фигурата съдържа броя на хората, демонстриращи всеки от режимите на обобщение (*обърнат*, с *базова съответсвие* и *смесен*) за експерименти: Експеримент 1: *ИБРЕ* с Учене чрез класификация и Експеримент 2: *ИБРЕ* с Учене чрез извод.

Междинна дискусия

Като цяло, от една страна, вербалните протоколи подкрепят очакването, че представянията на учещите чрез класификация наистина изграждат асиметрични представи за категориите. От друга страна, ученето чрез извод – за разлика от ученето чрез класификация – води до по-балансирано представяне на характеристиките на категориите. Въпреки това, *ИБРЕ* се наблюдава и при двата типа учещи, въпреки докладваните разлики в представите им. Като цяло тези резултати противоречат на твърдението, че асиметричното представяне е необходимо за наблюдаването на *ИБРЕ* (Johansen et al., 2007; Kruschke, 1996), тъй като учещите чрез извод докладват симетрични дефиниции на категории и все пак предпочитанията им следват обикновено наблюдаваните такива в контекста на *ИБРЕ*. Много по-вероятно е асиметричното представяне в класическата процедура (както предполага възгледа на Kruschke (1996), подкрепен от устно докладваните от участниците дефиниции на категориите) да е само страничен ефект от ученето чрез класификация. По-важното е, че емпиричните данни са както последователни, така и несъвместими с базираното на асоциации обяснение на ефекта (Kruschke, 1996). От една страна, по-честите категории действително се научават по-рано и участниците повече или по-малко докладват съдържанието, възприето от Kruschke (1996). От друга страна, въпреки че двете задачи се различават в своите изисквания за внимание и водят до различни репрезентации (както се вижда и от разликите в докладваните дефиниции), *ИБРЕ* се наблюдава и при двете задачи.

Интересното е, че резултатите са както в съответствие с базираното на правила обяснение на ефекта, така и несъвместими с него (Juslin et al., 2001). От една страна, обяснението, базирано на правила, очаква *ИБРЕ* и в двете условия на учене. От друга страна, това обяснение предполага подробни и симетрични представяния и при двата типа учене (както чрез класификация, така и чрез изводи), което противоречи на вербалните доклади на участниците, показващи повече асиметрия на представянията на учещите чрез класификация.

7. Експеримент 3: *ИБРЕ* с Мотивация преди ученето

Обосновка на Експеримент 3 и 4

Melchers et al. (2008) предлагат обширен преглед на емпирични данни, показващи, че манипулации като предварително обучение, инструкции към задачи, мотивация, и т.н., имат ефект върху стратегиите за кодиране, които участниците прилагат. Тъй като стратегиите за кодиране по време на ученето се променят, умствените представи за придобитата информация също са засегнати и по този начин по-нататъшното прилагане на тази информация от страна на участниците. Следователно, очакването е, че ако *ИБРЕ* наистина е ефект, дължащ се на процеси на ученето, допълнителна мотивация преди учебната фаза би следвало да модулира ефекта. Съответно третият и четвъртият експеримент тестват ефекта в условия на мотивацията преди ученето и преди тестването. В случай че магнитутът на *ИБРЕ* между тези два експеримента (и първия) се различава и в зависимост от посоката на ефекта, можем да правим допълнителни изводи дали ефектът е модулиран от процеси на ученето или от такива на тестване.

Участници

Общо 64 участници взеха участие в експеримента в замяна на частичен кредит за курса. Единадесет от тях бяха изключени от анализа поради ненадминаване на прага на учене от най-малко 70% правилни отговори както за честите, така и за редките категории в третия блок от фазата на учене. Така крайната извадка се състои от 53 участници (средна възраст = 29.3 години, SD = 9.9, 31 жени).

Материали

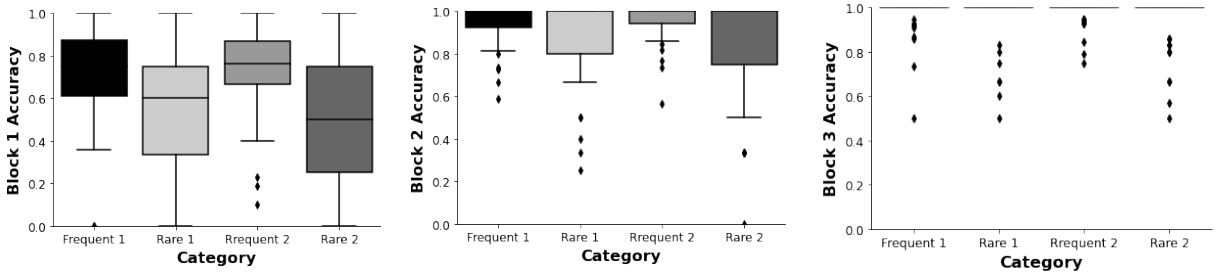
Стимулните материали имитираха тези в Експеримент 1: *ИБРЕ* с Учене чрез класификация (Фигура 2).

Процедура

В допълнение и за разлика от предишните експерименти, точно **преди фазата на учене** участниците бяха уведомени, че трябва да се опитат да се представят възможно най-добре, тъй като справилите се трима най-добре в експеримента ще получат ваучер за книжарница Orange на стойност от 50 лв. Във всички други аспекти експериментът се придържаше към фазите на учене и на тестване, описани по отношение на Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Резултати и Дискусия

Учене. Делът на верните отговори в първата част на ученето е по-висок за по-честите (.72), в сравнение с по-редките категории (.52) ($t(52) = 6.22$, $p < .0001$, $d = 0.854$ с 95% *CI* [0.14, 0.27]). Тази разлика намалява до края на третата и последна част от фазата на учене (0.98 за честите и 0.96 за редките категории), но остава значима – ($t(52) = 2.39$, $p = .02$, $d = 0.328$ и 95% *CI* [0.00, 0.04]). Като в Експеримент 1: *ИБРЕ* с Учене чрез класификация, често срещаните категории са научени много по-рано от редките (Фигура 12). Въпреки това до края на учебната фаза категориите са добре усвоени.



Фигура 12. Средна точност и ст. отклонение по тип категория в Блок 1, Блок 2 и Блок 3 отляво надясно за Експеримент 3: *ИБРЕ* с Мотивация преди ученето.

Тестване. Предпочитанията на участниците (Таблица 6) ясно показват тези, асоциирани с *ИБРЕ*. Хи-квадрат анализа (с 50:50 очаквания стойности), изчислен върху отговорите на участниците, показва следните предпочитания: при *Комбинираните* тестови случаи се наблюдава уклон към по-рядката категория – $\chi^2(1, N=195) = 17.85$, $p < .0001$, $\varphi = .3$ и 95% *CI* [55, 82] и [113, 140] съответно за по-честите и по-редките категории; за *Общите* тестове участниците имат силно предпочитание да отговарят с по-честата категория – $\chi^2(1, N=199) = 51.26$, $p < .0001$, $\varphi = .51$ и 95% *CI* [137, 162] и [37, 62] за по-честата и за по-рядката категория. За тестовите случаи от тип *Всички заедно*, хората демонстрираха леко, но незначимо предпочитание към базовата честота на категориите – $\chi^2(1, N=197) = 1.14$, $p = .285$, $\varphi = .07$, 95% *CI* [92, 120] и [77, 105] за по-честите и по-редките категории съответно.

Таблица 6: Съотношение на предпочитаните категории по тип тест за Експеримент 3: *ИБРЕ* с Мотивация преди ученето.

Test Cases	Choice proportion			
	Frequent	Rare	Frequent O	Rare O
Unique to F	**0.77**	**0.14**	**0.06**	**0.03**
Unique to R	**0.1**	**0.85**	**0.02**	**0.03**
Common	**0.71**	**0.23**	**0.03**	**0.03**
Combined	**0.32**	**0.6**	**0.03**	**0.05**
All together	**0.5**	**0.43**	**0.04**	**0.03**

8. Експеримент 4: *ИБРЕ* с Мотивация преди тестването

Участници

Общо 64 участници взеха участие в експеримента в замяна на частичен кредит за курса. Три от тях бяха изключени от анализа, тъй като не достигнаха прага на учене от най-малко 70% правилни отговори както на честите, така и на редките категории в третия и последен блок от фазата на учене. Финалната извадка се състоеше от 61 участници (средна възраст = 26.2 години, SD = 8.4, 43 жени).

Материали

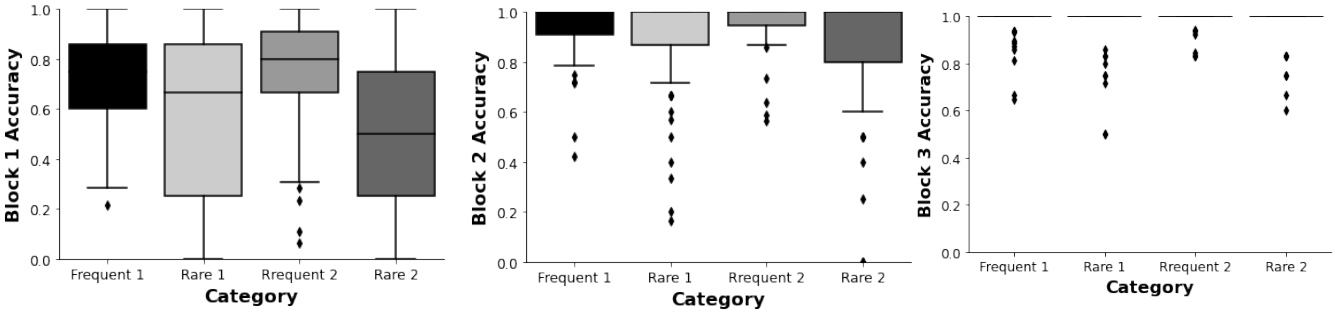
Стимулният материал беше идентичен с този в Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Процедура

Като в Експеримент 3: *ИБРЕ* с Мотивация преди ученето, експериментът се състоеше от две фази: фаза на учене и фаза на тестване. Важно за този експеримент е, че не бе предоставена каквато и да е мотивация преди ученето. Противно на Експеримент 3: *ИБРЕ* с Мотивация преди ученето, допълнителният паричен стимул беше предоставен точно **преди тестовата фаза**, където участниците бяха уведомени, че трябва да се постараят да се представят максимално добре, тъй като най-добре класиралите се 3-ма в експеримента ще получат ваучер за книжарница Orange на стойност 50 лв. Във всички други отношения фазите на обучение и тестване на експеримента имитираха тези от Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Резултати и Дискусия

Учене. Делът на верните отговори в първата част на ученето е по-висок за по-честите категории (.73), в сравнение с по-редките такива (.53) ($t(60) = 5.72$, $p < .0001$, $d = 0.732$ с 95% *CI* of [0.13, 0.27]). До края на третата и последна част от фазата на учене тази разлика изчезва – ($t(60) = 1.73$, $p = .089$, $d = 0.222$ с 95% *CI* от [-0.003, 0.035]). Като в Експеримент 1: *ИБРЕ* с Учене чрез класификация и Експеримент 2: *ИБРЕ* с Учене чрез извод, често срещаните категории са научени много по-рано от редките (Фигура 15). Въпреки това до края на учебните опити категориите са добре усвоени.



Фигура 15. Средна точност и ст. отклонение по тип категория в Блок 1, Блок 2 и Блок 3 отляво надясно за Експеримент 4: *ИБРЕ* с Мотивация преди тестването.

Тестване. Както е показано на Таблица 7, предпочитанията, свързани с *ИБРЕ*, се наблюдават и тук. Хи-квадрат анализът, изчислен върху честотите на предпочитания при *Комбинираните* тестови случаи, показва силно предпочитание на по-рядката категория – $\chi^2(1, N=234) = 36.17$, $p < .0001$, $\varphi = .393$ и 95% *CI* [57, 86] и [148, 177] за по-честите и по-редките категории. При *Общите* тестови случаи хората се придържаха силно към базовите честоти на категориите – $\chi^2(1, N=234) = 63.61$, $p < .0001$, $\varphi = .521$ и 95% *CI* [164, 191] и [44, 70] за по-честите и по-редките категории съответно. Наблюдава се и малък, но значим уклон – $\chi^2(1, N=237) = 7.8$, $p = .0052$, $\varphi = .181$ и 95% *CI* [125, 155] и [82, 113] респективно за по-честите и по-редките категории при тестовите случаи от тип *Всички заедно*.

Таблица 7: Съотношение на предпочитаните категории по тип тест за Експеримент 4: *ИБРЕ* с Мотивация преди тестването.

Test Cases	Choice proportion			
	Frequent	Rare	Frequent O	Rare O
Unique to F	**0.89**	**0.07**	**0.03**	**0.01**
Unique to R	**0.08**	**0.86**	**0.03**	**0.03**
Common	**0.72**	**0.23**	**0.02**	**0.03**
Combined	**0.29**	**0.69**	**0.01**	**0.01**
All together	**0.57**	**0.39**	**0.01**	**0.03**

ИБРЕ при различни условия на мотивация (без допълнителна мотивация, мотивация преди ученето, мотивация преди тестването)

Като цяло манипулирането на мотивацията в рамките на процедурата, асоциирана с *ИБРЕ*, (чрез допълнително мотивиране на участниците с ваучери преди фазата на учене в Експеримент 3: *ИБРЕ* с Мотивация преди учене и преди тестовата фазата в Експеримент 4: *ИБРЕ* с Мотивация преди тестване) не влияе на големината на ефекта, измерен чрез предпочитанията за класификация при *Комбинираните* тестови случаи, $\chi^2(1, N=428) = 1.13$, $p = .287$, $w = .05$. Интересното е, че има малка, но значима разлика в големината на ефекта между двата типа манипулация на мотивацията и Експеримент 1: *ИБРЕ* с Учене чрез класификация – $\chi^2(1, N=661) = 6.89$, $p = .032$, $w = .10$. Изглежда, че ефектът е по-силен с допълнителан мотивация. Тъй като не се наблюдава разлика между двата мотивиращи сценария, допълнителната мотивацията следва да има ефект най-вече по време на тестовата фаза (тъй като мотивацията, получена преди фазата на учене (т.е. Експеримент 3), присъства и по време на тестването, докато получената мотивация преди фазата на тестване (т.е. Експеримент 4) не включва допълнителна мотивация преди ученето). Следователно разликата в големината на ефекта (в сравнение с ефекта в Експеримент 1: *ИБРЕ* с Учене чрез класификация) може да се разглежда като вид подкрепа за подходи, приписващи *ИБРЕ* на някакъв тип рационални процеси.

9. Експеримент 5: *ИБРЕ без Учене* (*необходимо ли е учене за наблюдаване на ИБРЕ*)

Обосновка на Експеримент 5

Логичен следващ въпрос е дали *ИБРЕ* изобщо е феномен, движен от процеси, свързани с ученето, или може да възникне и в задача за вземане на решения. В опит да намерят минимално необходимите условия за наблюдаване на *ИБРЕ*, Johansen et al. (2007) правят опит да отговорят на този въпрос, демонстрирайки че задача за вземане на решение, при

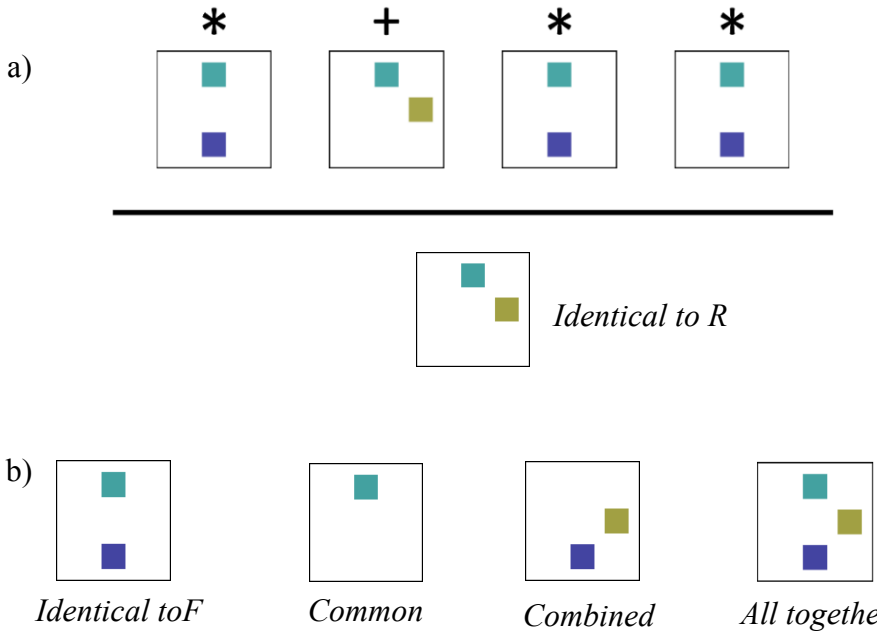
която честотната информация за категориите се предлага експлицитно в обобщен формат, не е достатъча, за да се наблюдава *ИБРЕ*, и правят извода, че за да се наблюдава ефекта, е необходимо пренебрегване на тази информация. Трябва да се вземе предвид, че Johansen et al. (2007) представи всички инструкции, примери на категории (вкл. експлицитна информация за честотите на категориите) и, най-важното, всички тестови опити на една и съща страница, позволявайки появата на специфични алтернативни стратегии, които вероятно не се появяват в класическия вариант на *ИБРЕ* (напр., изрични сравнения на разликите между тестовите примери). Дори по-важното е, че инструкциите към участниците включват фрази като „... *вие сте лекар в обучение*...“ и „... *след като сте прочели това внимателно*...“ (позовавайки се на примерите на категориите, видяни по време на ученето), като и двете намекват, че нещо трябва да се научи преди да се премине към тестовите примери), което отнема от първоначалната идея за задача за чисто вземане на решение.

Участници

Общо 75 участници взеха участие в експеримента в замяна на частичен кредит за курса. Дванадесет от тях бяха под критериите за не повече от 5 грешки при всички идентични типове контролни тестове (и не повече от 4 при един тип контролен тест). Така крайната извадка се състоя от 63 участници (средна възраст = 24.5 години, SD = 6.4, 52 жени).

Материали

Стимулните материали се състоят от четирите цветни квадрата от първите два експеримента (Фигура 2). Беше въведена и допълнителна променливост между опитите. Това беше реализирано чрез представяне на характеристиките 4 (а не в 2) възможни позиции за всеки пример на категория. Фигура 19, *а*) предлага визуализация на един опит.



Фигура 19. Панел а) показва един (*Идентичен на по-рядката категория*) опит в Експеримент 5: *ИБРЕ* без Учене. (*Бележка*. В действителнст етикетът "*Идентичен на по-рядката категория*" не присства на екрана). Панел б) показва всички топиве тестови стимули, които биха могли да се представят под линията.

Процедура

Във всеки опит от експеримента участниците бяха представени със стимул, подобен на този, представена на Фигура 19, *а*) (но без надписа "*Идентичен на по-рядката категория*"). Задачата на участниците беше да отговорят "*Какъв е стимулът под линията – '*' или '+'*". Отговорите бяха събрани чрез натискане на клавиш (В за '*' и М за '+'). Като в Експеримент 1: *ИБРЕ* с Учене чрез Класификация и Експеримент 2: *ИБРЕ* с Учене чрез извод, всички опити започваха с фиксационен кръст, представен за 500 ms и завършваха с 1000 ms междупробен интервал (ITI). След отговорите не беше предоставена коригираща обратна връзка. От изключителна важност е, че опитите не бяха независими един от друг и цялата инфромация, за вземане на решени за даден опит се съдържаше в самия опит. Всеки един опит беше създаван онлайн – с характеристиките (цветовете), позициите на цветовете, позицията на примерите от двете категории и т.н. генерирани на случаен принцип.

От изключително значение тук е, че тази експериментална обстановка позволява честотната информация за категориите (стимулите над хоризонталната линия) да бъдат представени едновременно и всеки опит да действа като самодостатъчен тестов случай, който е независим от останалите. С други думи, влиянието на ученето беше сведено до минимум (в действителност броят на идентичните опити в експеримента рядко надвишава два на участник). От една страна, тази така конструираната задача позволява включването на всички

тестови случаи, обикновено тествани в парадигмата на *ИБРЕ*. Фигура 19, *а*) представя един примерен опит (в този случай целевият пример, който се нуждае от класификация, е идентичен на една от категориите). Фигура 19, *б*) съдържа по-изчерпателен списък на критичните типове тестове. От друга страна, задачата позволява въвеждането на честотни разлики между примерите на категориите (представени над линията). Използвани са както съотношения между примерите 3:1 (като примера на Фигура 19, *а*), така и контролни случаи със съотношения 2:2, в които всяка от двете категории е представена от 2 примера. Очакванията бяха, че в контролните случаи на всеки от критичните опити участниците ще отговорят произволно. Експериментът се състоеше от по 100 опита на участник – за тестовите условия с честота 2:2 имаше по 5 опита за тип тест; за тестовото условие 3:1 имаше по 10 опита за *Идентичен на по-честата категория* и *Идентичен на по-рядката категория* тестови случи и по 20 за критичните типиве (*Общ*, *Комбниран* и *Всички заедно*). Причината зад тази разлика между броя на опитите в различните съотношения и типове тестове е изцяло практическа (така че експериментът да е с разумна продължителност).

Резултати и Дискусия

Тестване. Първо разгледан беше въпросът дали условието 2:2 наистина служи като контролно и няма определени предпочитания при някои от видовете тестове (с изключение на тестовете *Идентричен на контрола 1* и *Идентричен на контрола 2*). Очаквано, не бяха наблюдавани предпочитания към нито една от категориите: за *Общите* тестове *р*езултатите показват $\chi^2(1, N=315) = 0.03$, $p = .866$, $\varphi = .05$ с 95% *CI* от [138, 174] и [141, 177] за всяка една от категориите; предпочитания към дадена категория не бяха наблюдавани и по отношение на *Комбинираните* тестове ($\chi^2(1, N=315) = 0.92$, $p = .338$, $\varphi = .01$ с 95% *CI* от [148, 184] и [131, 167] за двете категории, нито за *Всички заедно* тестове ($\chi^2(1, N=315) = 1.68$, $p = .195$, $\varphi = .05$ с 95% *CI* от [128, 164] и [151, 187].

Таблица 8: Съотношение на предпочитаните категории според честота и тип тест в Експеримент 5: *ИБРЕ* без Учене.

Test Cases	**Choice proportion**			
	Frequent	**Rare**	**Control 1**	**Control 2**
Unique to F / Control 1	**0.97**	**0.03**	**0.95**	**0.05**
Unique to R / Control 2	**0.1**	**0.9**	**0.04**	**0.96**
Common	**0.74**	**0.26**	**0.5**	**0.5**
Combined	**0.54**	**0.46**	**0.53**	**0.47**
All together	**0.66**	**0.34**	**0.46**	**0.54**

Резултатите от по-голям интерес са тези, които идват от предпочитанията при наличието на честотни разлики между категориите 3:1. Хи-квадрат анализа по отношение на *Общите* тестови случаи показват силен уклон към по-честите категории – $\chi^2(1, N=1260) = 287.62$, $p < .0001$, $\varphi = .478$ и 95% *CI* от [899, 996] и [299, 361] за по-честите и по-редките предпочитания. Същото се наблюдава и при *Всички заедно* тестове – $\chi^2(1, N=1260) = 125.72$, $p < .0001$, $\varphi = .316$, и 95% *CI* от [651, 721] и [539, 609]. И накрая, в случаите на *Комбинираните* тестове участниците демонстрират леко, но все пак значимо предпочитание към базовите честоти – $\chi^2(1, N=1260) = 9.96$, $p = .002$, $\varphi = .089$, и 95% *CI* от [651, 721] и [539, 609]. За обобщени пропорции на предпочитанията вижте Таблица 8.

В класическата версия на *ИБРЕ*, когато става дума за *Комбинираните* тестови опити, има обръщане на предпочитанията; което означава, че хората предпочитат по-редките отговори. Въпреки липсата на пълно обръщане в този случай, предпочитанията на участниците ясно показват тенденциите, свързани с *ИБРЕ* – честотата на честите отговори е най-висока за *Общия* тип тест (до 74%), следвани от тест *Всички заедно* (66%) и *Комбинираните* тестови типове с най-ниски предпочитания към по-честата опция (54%). Нещо повече, хи-квадрат анализът показва, че има значима връзка между типа критичен тест и предпочитанията към определена категория при разлики в честотите със съотношение 3:1, $\chi^2(2, N=3780) = 105.28$, $p < .0001$, $w = .167$. С други думи, нещо кара участниците да изберат по-често срещаната категория повече, когато виждат *Общия* тип тест; и нещо ги накара да изберат по-рядко често срещаната категория, когато видяха *Комбиниран* тип тест. Разбира се, има възможност причината да не наблюдаваме пълната версия на предпочитанията тук (което включва обръщане на предпочитанията, когато става въпрос за *Комбинираните* тестове) да е процедурна, напр., необходимостта от по-отчетливи разлики по отношение на съотношенията между категориите (като 7:1). Shanks (1992) докладва *ИБРЕ* в неговата класическа парадигма със съотношения 7:1, но не и с 3:1.

10. Експеримент 6: *ИБРЕ* с Контролно условия (*нужна ли е честотна разлика за наблюдение на ИБРЕ*)

Обосновка на Експеримент 6

Привидно игнорирана подробност е, че *ефектът на обратната базова честота* рядко се тества с контролни условия, напр., когато и двете категории в двойката се появяват еднакъв

брой пъти. Доколкото ми е известно, ефектът никога не е бил тестван и наблюдаван в условия, при които структурата на категориите в контролните двойки е една и съща (всяка категория има обща и уникална характеристика), но няма разлики в честотата между категориите в двойката. Ако *ИБРЕ* се наблюдава за двойката категории с честотни разлики (но не и в двойката без честотни разлики), това би означавало, че ефектът не може да бъде приписан на замърсяващи характеристики, свързани със стимулите, тъй като единствената разлика между двете двойки категории би била честотна разлика вътре в самата двойка категории. Поради това настоящият експеримент имаше за цел да тества едно от критичните условия за наблюдаване на *ефекта на обратната базова честота* – а именно необходимостта от честа и рядка категория, които трябва да се научат. Освен това данните от този експеримент бяха подложени на допълнителен експлораторен анализ. По-конкретно, сред основните цели беше да се проучи дали *ИБРЕ* се проявява при участници, които не са успели да достигнат критериите за научаване. Обосновката тук е, че ако *ИБРЕ* наистина разчита на придобиването на асиметрични представяния, тогава участниците, които не успеят да научат задоволително категориите, няма да го покажат.

Участници

Общо 170 участници взеха участие в експеримента в замяна на частичен кредит за курса. Петдесет и пет участници бяха включени в основните анализи поради недостигане на критерия за учене (70% верни отговори в третата и последна част на ученето за всяка от категориите). Т.е. финалната извадка се състои от 115 участници (средна възраст = 26.72 години, SD = 9.42, 95 жени).

Материали

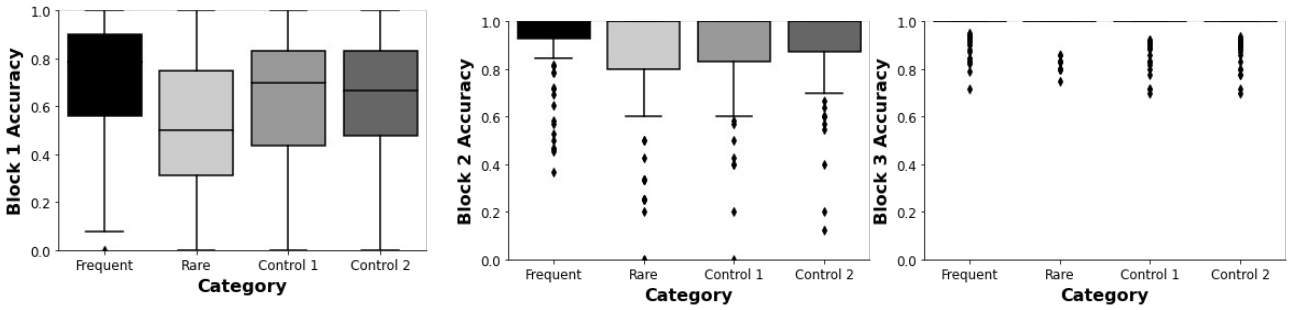
Характеристиките на стимулите са идентични с представените в Експеримент 1: *ИБРЕ* с Учене чрез класификация. Разликата се състоеше в структурата на една от двойките категории – докато една от двойките категории се състоеше от честа и рядка категория, категориите в другата двойка се появяваха еднакъв брой пъти – 30 опита за учене на категория.

Процедура

Експериментът имитира процедурата на Експеримент 1: *ИБРЕ* с Учене чрез класификация.

Резултати и Дискусия

Учене. Делът на правилните отговори в първите 40 учебни опита е различен за различните категории – ($F(3, 111) = 11.3$, $p < .001$, $\eta_p^2 = 0.069$) с обща средна успеваемост от 0.63 (по-конкретно, 0.72 за честата категория, 0.52 за рядката и по 0.63 за двете контролни категории). Пост-хок сравненията по двойки на Bonferroni показва, че разликата е само по отношение на рядката категория. В първата част на ученето участниците отговарят с повече грешки при по-редките в сравнение с по-честите примери (0.77) – $t(114) = 5.79$, $p = < .001$, $d = 0.763$ и 95% *CI* [0.11, 0.30]; и контролните примери – $t(114) = 3.38$, $p = .005$, $d = 0.446$ и 95% *CI* [0.03, 0.21] за една от контролите $t(114) = 3.32$, $p = < .006$, $d = 0.438$ и 95% *CI* [0.03, 0.21] за другата катеогрия в контролната двойка. Разликата в точността между категориите спада във втория блок ($F(3, 111) = 2.75$, $p < .042$, $\eta_p^2 = 0.018$). Последващи сравнения на Bonferroni по двойки показват, че единственото разлика идва от малка разлика между по-честата (0.93) и по-рядката категория (0.86) – $t(114) = 2.83$, $p = .03$, $d = 0.02$ и 95% *CI* [0.01, 0.13]. Както се очаква, честата категория е придобита много по-рано, последвана от контролната двойка и рядката, придобита най-късно (Фигура 22). До края на третата и последна част от фазата на учене тази разлика намалява напълно (0.98 както за честата, така и за рядката категория; 0.97 за двете контролни категории) – $F(3, 111) = 0.89$, $p < .449$, $\eta_p^2 = 0.006$. С други думи, до края на учебната фаза и четирите категории са усвоени еднакво добре.



Фигура 22. Средна точност и ст. отклонение за всяка категория в Блок 1, Блок 2 и Блок 3 отляво надясно в Експеримент 6: *ИБРЕ* с Контролно условие.

Тестване. Както се очаква (Таблица 9), когато става въпрос за двойката често-редки категории, хората правилно избират често срещаната категория, когато се представят с уникалната характеристика на по-честата категория (в 88% от случаите) и рядката категория, когато са представени само с уникалната ѝ характеристика (92% от случаите). Същите предпочитания се наблюдават и по отношение на контролната двойка (90% правилни класификации за една от контролните категории и 96% за другата).

Таблица 9: Съотношение на предпочитаните категории по тип тест за Експеримент 6: *ИБРЕ* с Контролно условие.

Test Cases	Choice proportion			
	Frequent	Rare	Control 1	Control 2
Unique to F / Control 1	**0.88**	**0.12**	**0.9**	**0.1**
Unique to R / Control 2	**0.08**	**0.92**	**0.04**	**0.96**
Common	**0.72**	**0.28**	**0.48**	**0.52**
Combined	**0.36**	**0.64**	**0.44**	**0.56**
All together	**0.58**	**0.42**	**0.53**	**0.47**

По-важното е, че предпочитанията по отношение на критичните тестови ясно показват поведението, свързано с *ИБРЕ* само в двойката категории с честотна разлика от 3:1. Хи-квадрат анализът, изчислен върху честотите на отговор за *Комбинираните* тестови случаи и очаквани стойности от 50:50, показват значителен уклон към рядката категория, $\chi^2(1, N=217) = 16.04$, $p < .001$, $\varphi = 0.27$ с 95% *CI* оотf [65, 94] за честата и [123, 152] за рядката категория. За *Общите* тестови случаи хората се придържат силно към базовите нива на категориите – $\chi^2(1, N=223) = 42.19$, $p < .001$, $\varphi = 0.44$ и 95% *CI* от [146, 173] и [50, 77] съответно за по-честата и по-рядката категория. И накрая, за *Всички заедно* тестови случаи хората демонстрират леко предпочитание към по-честата категория – $\chi^2(1, N=219) = 5.59$, $p = .018$, $\varphi = 0.16$ и 95% CI от [112, 142] и [78, 107] за по-честата и по-рядката категория. Както се предполага, тези уклони не се наблюдават, когато става въпрос за контролната двойка. Нито един от хи-квадратите в контролното условие не показа значително предпочитание към някоя от категориите – $\chi^2(1, N=218) = 0.29$, $p = .588$, $\varphi = 0.04$ (за *Общите* случаи); $\chi^2(1, N=224) = 3.5$, $p = .061$, $\varphi = 0.13$ (за *Комбинираните* тестови) и $\chi^2(1, N=212) = .68$, $p = .41$, $\varphi = 0.06$ (за *Всички заедно* тестове). С други думи, не е наблюдаван *ефект на обърнатата базова честота* за контролните категории. Само от това можем да заключим поне две неща: 1) че честотната разлика между категориите във фазата на учене действително е критично условие, за да се наблюдава *ИБРЕ*; 2) полученият *ИБРЕ* не може да се дължи на замърсяващи характеристики на стимулите, тъй като единствената разлика между двете двойки категории е честотите между категориите, образуващи всяка двойка.

Допълнителни Експлораторни анализи

За да се провери възможността *ИБРЕ* да се дължи на страничен ефект от някакъв вид уклон, свързан със самите отговори, беше изследвано съотношението между честите и редките избори в цялата тестова фаза (т.е. колко пъти е избран всеки един от отговорите/е натиснат

бутони). Изглежда хората са балансирани в избора си – по-чест отговор е даден на 51,78% от целевите опити в сравнение с 48,22% за по-редките.

Таблица 11: Съотношение на предпочитаните категории по тип тест за Експеримент 6: *ИБРЕ* с Контролно условие за участниците, които не са преминали прага на учене.

Test Cases	Choice proportion	
	Frequent	Rare
Unique to F	**0.74**	**0.26**
Unique to R	**0.4**	**0.6**
Common	**0.69**	**0.31**
Combined	**0.42**	**0.58**
All together	**0.55**	**0.45**

Направен е и проучвателен анализ на предпочитанията на участниците, които не са покрили прага на учене, поради което не са били включени в анализа по-горе. Техният избор за критичните тестове (Таблица 11) показва предпочитания, много тясно свързани с *ИБРЕ* – честотата на честите отговори е най-висока за *Общия* тип тест (до 69%), последвано от типа тест *Всички заедно* (55%) и типа *Комбиниран* тест с най-малко предпочитание към по-честите отговори (42%). Предпочитанието на по-честата категория при *Общите* тестови случаи показва силно предпочитание към по-честата категория – $\chi^2(1, N=91) = 13.46$, $p = < .001$, $\varphi = 0.62$. Въпреки че числово се наблюдават предпочитанията, свързани с *ИБРЕ*, нито предпочитанието на по-рядката категория при *Комбинираните* тестове, нито предпочитанието на по-честата категория при тестовете от тип *Всички заедно* показват статистическа значима разлика (съответно, $\chi^2(1, N=89) = 2.53$, $p = .112$, $\varphi = 0.17$ и $\chi^2(1, N=98) = 1.02$, $p = .117$, $\varphi = 0.10$). И все пак има изследователи, които не докладват резултати от статистически анализи, а третират цифровите предпочитания на участниците като достатъчно добра демонстрация на ефекта (напр., Lamberts & Kent, 2007). Липсата на значими разлики в предпочитанията при тестовете от тип *Всички заедно* и *Комбинирани* не е толкова изненадващо, тъй като чувствителността на последните две сравнения е под 0.4 (с други думи, броят на наблюденията е доста ограничен). Следователно не могат да се направят големи заключения само от този анализ. Въпреки липсата на статистическа значимост, числените предпочитания, свързани с *ИБРЕ*, са очевидни. Сам по себе си този резултат поставя под въпрос твърдения от типа, че *ИБРЕ* е ефект, дължащ се на процеси по време на ученето. Имайки предвид и разликата между различните видове режими на прилагане на наученото (*обърнат, в съответствие с честотата* и *смесен*), има смисъл потенциалните разлики между, хората при които се наблюдава ефекта и тези, при които не се наблюдава, да се търсят някъде другаде.

11. *ИБРЕ* с Езиков модел от тип трансформатор

И двата модела, които предлагат обяснение на *ИБРЕ* – *EXIT* (Kruschke, 2001) и *ELMO* (Juslin et al., 2001) – страдат от липса на генерализируемост. Те съдържат специфични механизми (механизми за учене, водещи до асиметрични представяния в случая на *EXIT* и изводи чрез елиминиране в случая на *ELMO*), предназначени да адресират конкретно *ИБРЕ*.

Едно изключение от тази тенденция, когато става въпрос за *ИБРЕ*, е моделът *RoleMap* (Petkov & Petrova, 2019). *RoleMap* е базиран на архитектурата с общо предназначение *DUAL* (Kokinov, 1988, 1994), която разглежда ефекта като резултат на релаксация на мрежа за удовлетворяване на ограничения на връзките, изразяващи два натиска с общо предназначение – тенденцията да се виждаме подобни неща като съответстващи и тенденцията да намираме едно нещо като съответстващо само на едно друго нещо.

Неизследвана посока е дали *ИБРЕ* все пак може да се наблюдава с архитектура с общо предназначение, базирана на асоциации. Езиковите модели от тип Трансформатор (TLMs) (Vaswani et al., 2017) са една група такива модели. Спецификата на тези модели се състои в техните представяния, които са изключително сложни и могат да бъдат адаптирани към голям брой задачи. Това е възможно благодарение на дизайна на архитектурата – тя е адаптирана да идентифицира значимостта на информацията, независимо от нейното местоположение. С други думи, той обработва корелации между далечни един от друг елементи във входния текст, като същевременно може да обръща внимание на някои думи повече от други. Като цяло, TLM са модели на статистическо разпределение на думите, извлечени от огромен корпус от естествени езикови текстови данни, т.е. генерирани от хора текстове. Те са генеративни, защото позволяват от тях да бъдат взимани проби, т.е. хора могат да „задават“ въпроси (чрез представяне на някакъв текстов фрагмент на входа на модела), а моделите могат да „отговорят“, като продължат текстовия фрагмент с думи, които е най-вероятно да бъдат следващи (Shanahan, 2022).

С появата на т.нар. *Generative Pre-trained Transformer 3 model* (*GPT-3)* – ярък пример на тази група модели (Brown et al., 2020) – беше демонстрирано, че TLM могат да имитират хората в едно конкретно отношение – те също могат да учат само от няколко примера (чрез т.нар. Базирано на подсказки инженерство, Zhang et al., 2021). Съществената разлика в базирания на подсказки подход е именно, че успява да се научи да изпълнява задача само с няколко примерла (без нуждата от хиляди примери, от каквито се нуждаят архитектурите на

предшестващите модели). Именно този подход е използван за симулацията, представена по-долу.

11.1. Симулация: ИБРЕ с GPT-3

Обосновка на Симулацията. Като цяло *GPT-3* не прави нищо повече от извличане на сложни статистически закономерности от огромно количество писмен естествен език. Важно е, че моделът се адаптира към изпълнение на задача без да актуализира асоциациите си (т.е. без никакви промяна на представянията си). Базираната на подсказка процедура се разглежда повече като обуславяне, а не като учене чрез формиране на нови представяния и промяната им (Brown et al., 2020; Radford et al., 2019). Това позволява лесно разграничаване между замразеното състояние на модел от промените, дължащи се на учене. По този начин, поради своята специфика, моделът може да действа като тестов инструмент, с който да се провери дали *ИБРЕ* разчита на процеси, протичащи по време на ученето (както твърди Kruschke, 1996) или трябва да бъдат обмислени друг тип механизми. Ако подобни на *ИБРЕ* „предпочитания“ се наблюдават в *GPT-3*, това би било ясна демонстрация на поне две възможности: 1) процеси, базирани на асоциации, са достатъчни, за да се появи *ИБРЕ* (тъй като *TLM* се основават на подобни асоциациативни статистически разбирания на езика); 2) *ИБРЕ* не е просто ефект, управляван от процеси, свързани с ученето, който разчита на асиметрични представяния (тъй като подходът, базиран на подсказки, не променя представянията на модела).

Материали. За тази цел беше използвана добре установена в литературата, свързана с *ИБРЕ,* версия на стимулния материал. Стимули следваха тези, използвани от Kruschke (Kruschke, 1996). По-конкретно, използвани бяха 6 думи/фрази, третирани като характеристики на категориите – а именно *болки в ушите*, *кожен обрив*, *болки в гърба*, *замайване*, *болки в мускулите* и *запушен нос*. Характеристиките бяха представени на модела вербално в писмен формат. Две категории с припокриващи се характеристики бяха проектирани за всеки симулационен цикъл. Една от категориите беше по-честа от другата. Двете категории бяха обозначени произволно с 2 латински букви: „*F*“ и „*R*“. Всяка от двете категории беше определена от две характеристики – една от характеристиките беше уникална за категорията, а другата беше споделена между двете категории.

Процедура. Един симулационен цикъл се състоя от комбинация от 60 категоризирани примера с честотна разлика 3:1 (45 примера от често срещаната категория и 15 примера от рядката) със съответните им етикети и един некатегоризиран тестов пример, за който се изискваше отговор. Симулацията бе направена в общо 300 цикъла (по 60 за всеки тестов тип

– *Уникална за F*, *Уникална за R*, *Общ*, *Комбиниран* и *Всички заедно*). За всеки отделен симулационен цикъл бяха конструирани уникални произволни комбинации за *двете категории* (т.е. кои три от шестте възможни характеристики ще бъдат включени в симулационния цикъл); *разпределението на характеристиките по категории* (коя ще е общата характеристика, коя ще бъде уникалната характеристика за честата категория и коя ще бъде уникалната характеристика за рядката категория); *поредността на примерите* (така че честите и редките примери да са предоставени в смесена поредност); и *позициите на характеристиките една спрямо друга* (стремейки се към произволно пространствено разпределение, така че общата характеристика да бъде представена относително равен брой пъти от лявата и от дясната страна спрямо уникалните характеристики). Опитите бяха администрирани като подкани на естествен език и беше записвано продължаването на текстовата поредица, което моделът предоставя. Всеки отговор беше класифициран като предпочитание към по-честата или по-рядката категория (както се прави с човешките данни).

Резултати и дискусия. Както е резюмирано в Таблица 12, когато е представен със сюжет, имитиращ парадигмата на *ИБРЕ*, моделът демонстрира подобни на *ИБРЕ* предпочитания, които наблюдаваме при хората. По-конкретно, когато е представен с уникална характеристика, моделът реагира правилно (той избира с голяма сигурност често срещаната категория, когато се представя с уникалната за честата категория характеристика и обратното за уникалната характеристика на рядката категория). Моделът предпочита по-честата опция, когато е представен с *Общия* тестов случаи и обръща предпочитанието си, когато е представен с *Комбиниран* тест. Предпочитанията, демонстрирани при тестовете *Всички заедно*, са някъде по средата.

Таблица 12: Съотношение на предпочитаните категории за тип тест за Симуалция: *ИБРЕ* с *GPT-3* модел.

Test Cases	**Choice proportion**	
	Frequent	**Rare**
Unique to F	**0.9**	**0.1**
Unique to R	**0.3**	**0.97**
Common	**0.67**	**0.33**
Combined	**0.4**	**0.6**
All together	**0.5**	**0.5**

Важно е, че резултатите не се дължат на реда на характеристиките, в които са представени по време на тестването (т.е. дали първата представена характеристика е уникалната за честата или за рядката категория) нито каквато и да е друга предходна вероятност на стимулите (т.е. възможността някои от характеристиките да са по-разпространени в естествения език и

поради тази причина моделът да им обръща повече внимание). Това заключение може да се направи от вероятностния спектър на токените[1], които могат да се разглеждат като по-подробна мярка както за посоката (т.е. дали изборът е честата или рядката категория), така и за силата на предпочитанието. Отнася се до очакването на модела да „види“ точно думите, с които е представен, и вероятността да продължи последователността от думи по някакъв специфичен начин (т.е. предпочитанието за категоризиране). Например при случаите на *Комбиниран* текст моделът демонстрира същото рядко предпочитание – както когато тестовият случай започва първо с уникалната характеристика на честата категория, така и когато тестовият случай започва с уникалната характеристика на рядката категория (0.5934 в първия случаи и 0.6071 във втория). Всъщност вероятностите за генерализиране (вероятността за класифициране на примера като пример на честата или рядката категория) са много по-повлияни от реда, в който са подадаени примерите на категориите. Въпреки че моделът има по-високи предпочитания за рядката категория за всяка една от изследваните поредици от примери, изглежда, че редът на примерите влияе върху силата на предпочитанията със сложен модел на предпочитания – от една страна изглежда, че има предпочитание към редуващи се отговори; от друга, тази пристрастност изглежда отслабва, когато има повторение в примерите.

Трудно е еднозначно да се каже, че този резултат пряко подкрепя асоциативния или базирания на правила подход. Базираният на правила подход предполага, че често срещаната категория е по-добре научена и по-скоро игнорира реда на примерите, представени във фазата на учене (Juslin et al., 2001). Резултатът по-скоро противоречи на подхода, базиран на асоциациите, представен от Kruschke (1996), тъй като предположението за асиметрично представяне разчита на това, че по-честата категория се придобива първа (тъй като от самото начало участниците виждат повече примери от тази категория).

11.2. Междинна дискусия

Като цяло, моделът *GPT-3* (и моделите, подобни на него) успешно възпроизвеждат човешки данни, свързнаи с проследяване на очите, времена за четене и други психологически феномени (Merkx & Frank, 2021; Schrimpf et al., 2020; Marinova et al., 2021). Както вече е отбелязвано, *GPT-3* изпълнява на човешко ниво редица НЛП задачи, поради което е и сред най-разпространените трансформиращи модели, изучавани от когнитивните психолози (Binz

1 Тъй като TLM имат стабилни представяния, *GPT-3* и неговите вероятностни спектри могат допълнително да се използват като инструмент за изследване относно това какво може да ръководи ефекта. Вероятностният спектър на модела съдържа информация за първите пет потенциални завършвания на конкретна подкана и свързаните с тях вероятности.

& Schulz, 2022). Тъй като е обучен с текстови данни, създадени от човека, се очаква да е кодирал различни уклони, появаващи се у хората – напр., демонстрира различни пристрастия, свързнаи с пола, и редица уклони, когато е подканван да генерира истории (Lucy & Bamman, 2021); предлага различни професии в зависимост от пол, раса и сексуална ориентация (Sheng et al., 2020). Например, проявява същите евристики и пристрастия като хората, когато им се представят класически проблеми като „проблемът Линда“ и „болничен проблем“ (Binz & Schulz, 2022). Въпреки това, този тип модели със сигурност не притежават способности за разсъждение на високо ниво. Например *GPT-3* има сериозни затруднения със задачите за извод (напр., т.нар. *ANLI* набор от данни и задачи, Brown et al., 2020), и не показва признаци на насочено изследване, което е много характерно за хората (Binz & Schulz, 2022).

Фактът, че подобни на *ИБРЕ* предпочитания се наблюдават с модел като *GPT-3* може да се счита за подкрепа на идеята, че процесите на разсъждение от по-високо ниво не са критични за постигане на ефекта. *GPT-3* е нищо друго освен статистически инструмент, обучен да предсказва следващата(ите) дума(и), при подадена поредица от думи. Той работи по чисто асоциативен начин, разчитайки на вероятностни разпределения на поредици от думи и представяния, които са просто статистически закономерности. Следователно резултатите от симулацията поставят под въпрос и двата възгледа – възгледи, приписващи *ИБРЕ* на процесите на разсъждение на високо ниво, и обяснението на Kruschke (1996) за *ИБРЕ* като резултат от придобито по време на ученето асиметрично представяне – и предполагат, че ефектът може да се обясни с друга статистическа закономерност, която *GPT-3* може да улови, като натиск за равномерно разпределение на наличните отговори.

Дискусия и Заключения

Вече повече от 30 години феномен, наречен ефект на обърнатата базова честота, оказва натиск за обяснение върху литературата за категоризиране. Да повторим, *ИБРЕ* се свързва с предпочитание за приписване на конкретни двусмислени примери на по-малко разпространени категории. Обикновено това предпочитание се появява заедно с предпочитание към по-честата категория, когато човек е представен единствено със споделена между двете категории характеристика. Както Don et al. (2021) отбелязват, изследването на генерализируемостта на ефекта е дълбоко пренебрегната. Дълго време подобните на *ИБРЕ* предпочитания се приписват на конструкции, свързани с вниманието и асоциациите. По-конкретно, ефектът се разглежда като резултат от придобитите асиметрични представяния по време на ученето на категориите (Kruschke, 1996, 2009). Въпросът дали

придобитата представна асиметрия не е критично условие за ефекта, а просто страничен ефект от самото учене чрез класификация (т.е. не е причината, поради която се появява *ИБРЕ*, а по-скоро паралелна характеристика на човешкото поведение след учене чрез класификация) никога не е било изследвано.

Сред целите на тази дисертация беше да продължи дебата относно механизмите, които са в основата на *ИБРЕ*. По-конкретно, тезата имаше за цел да тества предполагаемата роля на заучените асиметрични представяния за наблюдаване на ефекта (Експерименти 1 до 3) и дали ученето изобщо е необходимо (Експерименти 4 до 6). В допълнение чрез симулация с вероятностен модел, базиран на архитектура тип трансформатор, беше допълнително тествано дали ученето е необходима предпоставка за ефекта или може да бъде наблюдаван без промени в репрезентацията/ученето или друго.

Свързване на резултатите от експерименталните постановки с обясненията на ИБРЕ, базирани на асоциации и правила

В шест експеримента силно подкрепяното обяснение – че *асиметричните представяния*, усвоени при ученето стоят зад *ИБРЕ* (Kruschke, 1996, 2009) – беше поставено на тест и беше повдигната възможността ефектът да бъде поне частично модулиран и от други процеси. Първият експеримент (Експеримент 1: *ИБРЕ* с Учене чрез класификация) докладва репликация на класическата парадигма на *ИБРЕ* и служи като уверение, че ефектът може да бъде наблюдаван с нови визуални материали и разлика в съотношението между категориите 3:1. Вторият експеримент (Експеримент 2: *ИБРЕ* с Учене чрез извод) въведе важна процедурна разлика – обикновено използваната задача за учене чрез класификация (която може да доведе до представна асиметрия като страничен ефект) беше заменена с учене чрез извод (с очакването, че задачата ще попречи на формирането на такива представени асиметрии). Резултатите от този експеримент оспорват базираното на асоциация обяснение, тъй като въвежда постановка, която възпрепятства представната асиметрия и въпреки това *ИБРЕ* се наблюдава. Резултатите от Експеримент 1: *ИБРЕ* с Учене чрез класификация и Експеримент 2: *ИБРЕ* с Учене чрез извод и потенциалните разлики в представянията, до които водят двете задачи, се подкрепят от устните доклади на участниците. Според тях участниците наистина формират различни представяния на категории – учещите чрез класификация докладват дефинициите на категориите като асиметрични, докато учещите чрез изводи са по-склонни да ги докладват като дефинирани и от двете им характеристики.

Изглежда *ИБРЕ* се наблюдава въпреки разликите в усвоените представяния и липсата на асиметрични представяния при категориите, споделящи припокриваща се характеристика.

В допълнение, третият и четвъртият експеримент (Експеримент 3: *ИБРЕ* с Мотивация преди ученето и Експеримент 4: *ИБРЕ* с Мотивация преди тестването) тестваха ефекта в условия с допълнителна мотивация, предлагащи стимул преди фазата на учене и преди фазата на тестване. Въпреки че допълнителната мотивация засилва ефекта и магнитута му е по-висок в сравнение с класическата версия на парадигмата (Експеримент 1: *ИБРЕ* с Учене чрез класификация), не се различава в зависимост от това кога е администриран допълнителният паричен стимул, т.е. дали преди ученето (Експеримент 3: *ИБРЕ* с Мотивация преди ученето) или преди тестването (Експеримент 4: *ИБРЕ* с Мотивация преди тестването). Това води до извода, че ако не се управлява, то *ИБРЕ* е поне модулиран от базирана на правила обработка, появяваща се по време на тестването, тъй като мотивацията, получена преди фазата на ученето, все още може да има ефект върху тестовата фаза (реално мотивацията по време на тестването е общият фактор между двата експеримента).

За петия експеримент беше разработена задача за вземане на решения, при която ученето беше напълно елиминирано. Използваната експериментална обстановка позволи честотната информация, свързана с категориите, да бъде представени едновременно и всеки опит да действа като самодостатъчен тестов случай, който е независим от останалите. И все пак предпочитанията, свързани с *ИБРЕ*, бяха наблюдавани. Тези резултати поставят под съмнение всички възгледи, виждащи ефекта като следствие от ученето. Резултатите са в много по-голямо съответствие с обясненията на ефекта, които разчитат на разсъждения върху правила и/или примери.

Шестият експеримент проверява *ИБРЕ* в контекста на контролно условие, при което при част от категориите нямаше честотни разлики. Тъй като ефектът се наблюдава само за категориите с честотни разлики (но не и при категориите без честотни разлики), ясно е, че не можем да припишем ефекта на други специфики на стимулите. По-скоро честотната разлика изглежда като необходимо условие за наблюдаване на ефекта. В допълнение, резултатите от този експеримент бяха проучени по-подробно. Едно от по-значимите открития в това отношение показа, че „неуспешните“ учащи – които обикновено се отстраняват от анализа на данните поради неуспех да надминат предварително зададените критерии за учене – всъщност показват същите предпочитания, които се асоциират с *ИБРЕ*. Следователно нито научената представна асиметрия, нито ученето като цяло изглежда са от решаващо значение за появата

на *ИБРЕ*. По-скоро *ИБРЕ* разчита на някакво базирано на примери разсъждение, което може да се случва по време на фазата на тестване (т.е. Експеримент 5: *ИБРЕ* без Учене) или базирано на учене с примери (Експерименти 2, 3, 4 и 6), към които може да се реферира по време на теста.

Докладваните резултати са разширени със симулация, базирана на асоциативна архитектура от тип трансформатор (по-конкретно, *GPT-3*). От една страна, забележимо е, че *ИБРЕ* се появява с такава архитектура, разчитаща на нищо повече, освен на статистическа корелация в естествения език. От друга страна, важно е да се отбележи, че ефектът е получен без каквато и да е промяна в представите на модела, което е аргумент срещу всички обяснения на *ИБРЕ*, които го приписват на процеси, свързани с ученето. По-скоро симулацията и по-конкретно изследваният вероятностен спектър на думите и реда на примерите в докладваните симулации сочат към идеята, че *ИБРЕ* може да се дължи на взаимодействие между два натиска – взаимодействие между това колко дълго пример на даден категория не се е появявал и някаква готовност да разпределим отговорите си 50:50 между възможните опции.

Като цяло, резултатите подкопават специфичните варианти както на наличните обяснения, базираните на асоциации (Kruschke, 1996), така и на обясненията, базираните на правила (Juslin et al., 2001). Доминиращото обяснение, базирано на асоциации (Kruschke, 1996), не може да обясни докладваните данни от първите три експеримента, тъй като разчита на асиметрични представяния, формирани по време на ученето на категориите. Резултатите не могат да се тълкуват прибързано като подкрепящи и алтернативното обяснение, основано на правила, отдавайки ефекта на инференциални процеси на разсъждение (Juslin et al., 2001). Поне този специфичен пример на базирано на правила обяснение на ефекта среща някои трудности при обяснението на резултатите от петия експеримент (при която фазата на учене е напълно премахната), тъй като той разчита на базирани на правила представяния, формирани също чрез фазата на учене, и предполага че *ИБРЕ* се дължи на ограничения на капацитета на паметта по време на вземането на решение. В контекста на петия експеримент (когато всички категории са представени на екрана и по този начин активни), специфичното базирано на правила обяснение предсказва случаен избор между двете категории (с други думи, прогнозира липса на ефект), какъвто – както беше демонстрирано – не е случаят.

Недостатъци на изследването

Резултатите от тезата оспорват доминиращите обяснения на *ИБРЕ* (т.е. обяснението, базиран на асоциации (Kruschke, 1996) и обяснението, базиран на правила (Juslin et al., 2001)).

Въпреки това тезата не предлага изчерпателно систематично изследване на каквито и да е специфични алтернативни механизми, които потенциално могат да лежат в основата на ефекта на обърнатата базова честота.

Финални заключения

Основните заключения от докладваните експерименти и симулации са, че нито репрезентативната асиметрия, придобита чрез учене (Експеримент 2: *ИБРЕ* с Учене чрез извод), нито самото учене (Експеримент 5: *ИБРЕ* без Учене) са критични за наблюдението на *ИБРЕ*. *ИБРЕ* изглежда устойчив при различни задачи: същото обръщане на предпочитанията беше установено за критичните тестови опити, независимо от задачата за учене. Следователно *ИБРЕ* не може просто да се третира само като страничен ефект от обикновено прилаганата задача за учене чрез класификация. Той по-скоро казва нещо за категоризирането на обекти с припокриващи се характеристики, които могат да бъдат общи за различни задачи (класификация срещу обучение чрез извод) и ситуации (учене срещу не-учене).

Като се има предвид, че *ИБРЕ* се появява както след учене, така и в ситуация без условия за учене, поне част от ефекта трябва да идва от процеси (или/и стратегии), генерирани по време на фазата на тестване. Всякакви обяснения, които отдават ефекта единствено на процеси, случващи се по време на ученето, ще имат трудности при обяснението на резултатите от Експеримент 5 (Експеримент 5: *ИБРЕ* без Учене).

Приноси на тезата

Методологични приноси

1) Тезата изследва *ефекта на обърнатата базова честота* по систематичен начин в шест експеримента, манипулиращи различни фактори (задачата за учене, мотивиращи стимули, сценарии за вземане на решения и контролни условия), като същевременно поддържа еднакви стимулни материали и тестова процедура. Това дава възможност за по-ясна представа за възможните детерминанти на ефекта и за по-добро отчитане на критичните за него условия, тъй като всички други променливи са константни.

2) Резултатите от шестте експеримента се анализират по последователен начин, което позволява сравнения между експериментите. Следователно наблюдаваните разлики не могат да се обяснят с различно изрязване на данни, различни критерии за ефективно учене или различни анализи.

3) Използвани са както качествени, така и количествени данни, за да се тества базираната на асоциация обяснение на *ИБРЕ* и по-конкретно твърдението, че ефектът се дължи на представна асиметрия. Това е важно, тъй като дава по-пълно описание на поведението на хората.

Емпирични приноси

1) *Ефектът на обърнатата базова честота* беше наблюдаван за първи път чрез задача за учене чрез извод. Изглежда, че задачи за учене, намаляващи репрезентативните асиметрии на учените категории, все пак водят до *ИБРЕ*. Това е важно, тъй като оспорва базираното на асоциации обяснение на ефекта (т.е. приписването на ефекта на асиметрични представяния). В допълнение, това е и тест за обобщаемостта на ефекта.

2) Дипломната работа представя за първи път косвено измерване на представите на учените категории, докладвани от участниците. Устните доклади разкриват, че действително учещите чрез класификацията придобиват асиметрични представяния за разлика от учещите чрез извод. И все пак *ИБРЕ* се появява и при двата вида задачи за учене.

3) Предпочитания, свързани с *ИБРЕ*, са получени в чисти условия, базирани на примери, без каквото и да е учене. Това е важно, тъй като поставя под съмнение твърденията, че ефектът е ефект на ученето (и по-специално, че се дължи на специфични представи, придобити по време на фазата на учене).

4) Беше демонстрирано, че *ИБРЕ* се появява и при участници, които по принцип се отхвърлят като лоши учещи. Това е важно, тъй като допълнително поставя под въпрос твърдения, свързващи ефекта с ефективно учене.

5) *ИБРЕ* беше тестван за първи път и в контекста на допълнителни мотивация. Тествано е паричното влияние, представено преди ученето и преди тестването.

6) За първи път *ИБРЕ* беше тестван в контекста на ново контролно състояние – ефектът не беше получен в контролно условие, характеризиращо се с липса на честотни разлики между категориите.

7) За първи път беше тествано дали *ефектът на обърнатата базова честота* може да бъде получен с архитектура, с общо предназначение, базирана на асоциации. Беше демонстрирано, че ефектът възниква в такъв модел (по-конкретно, *GPT-3*) без никакви промени в представянията/без учене. Това е важно, тъй като остава в противоречие с

доминиращия в момента възглед, според който, за да се наблюдава ефектът, е необходимо учене (и по-конкретно, придобиване на асиметрични представяния на целевите категории).

Теоретични приноси

Тезата представя резултати, които са теоретично предизвикателство и за двете обяснения на *ИБРЕ* – асоциативното и базираното на правилата.

Публикации, свързани с дисертацията

Marinova, I., **Petrova, Y**., Slavcheva, M., Osenova, P., Radev, I., & Simov, K. (2021). Monitoring Fact Preservation, Grammatical Consistency and Ethical Behavior of Abstractive Summarization Neural Models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 901-909).

Petkov, G., & **Petrova, Y.** (2019). Relation-based categorization and category learning as a result from structural alignment. The RoleMap model. *Frontiers in Psychology*, *10*, 563.

Petrova, Y., & Petkov, G. (2018). Role-Governed Categorization and Category Learning as a Result from Structural Alignment: The RoleMap Model. *International Journal of Computer and Information Engineering*, *12*(8), 578-585.