



PEDAGOGICKÁ
FAKULTA
Univerzita Karlova

Říjen 2023

Přidaná hodnota učitele

Rešerše zahraniční literatury

Yvona Kostecká



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



PEDAGOGICKÁ
FAKULTA
Univerzita Karlova

Důvody nedostatku učitelů

Pedagogická fakulta Univerzity Karlovy je od prosince 2020 řešitelem projektu OP VVV (CZ.02.3.68/0.0/0.0/19_076/0016404) s názvem Důvody nenastoupení absolventů pedagogických fakult do profese učitele, který reaguje na problematiku nedostatku učitelů v českém školství.

Cílem projektu je identifikovat zda, do jaké míry a proč absolventi pedagogických fakult a učitelských studijních programů na ostatních fakultách nenastupují k výkonu povolání, na které byli v průběhu svého vzdělávání připravováni. Nebo zda, do jaké míry a proč z profese učitele brzy odcházejí. Zároveň je ovšem cílem projektu identifikovat významné bariéry a faktory, které k nedostatku učitelů vedou, a to na různých úrovních:

- a) na individuální úrovni samotných učitelů či studentů připravujících se na profesní dráhu učitele;
- b) na úrovni vysokých škol a fakult, které studenty ke vstupu do učitelské profese připravují;
- c) na úrovni (především) základních a středních škol jako zaměstnavatelů učitelů;
- d) na úrovni státu zodpovědného (prostřednictvím ministerstva) za zajištění učitelů a péči o ně;
- e) na úrovni společenských předpokladů a podmínek pro výkon učitelského povolání.

Na základě analýz primárních i sekundárních zdrojů informací a dat z kvalitativních a kvantitativních šetření, rešerší zahraniční a domácí literatury a rozborů vybraných vzdělávacích politik budou v rámci projektu navrženy krátkodobé kroky i dlouhodobější opatření a strategie, které:

- 1) vytvoří předpoklady pro to, aby mezi poptávkou a nabídkou učitelů a pedagogických pracovníků na pracovním trhu v budoucnu již nenastávaly podobné výkyvy a nerovnováhy;
- 2) omezí nenastupování příslušných absolventů do povolání učitele a případně přispějí k návratu části těch, kteří již působí v jiných odvětvích a povoláních.

Přidaná hodnota učitele

Rešerše zahraniční literatury

Yvona Kostecká

Středisko vzdělávací politiky, Pedagogická fakulta Univerzity Karlovy

Říjen 2023

Obsah

I. Zjišťování přidané hodnoty ve vzdělávání	6
I.1 Přidaná hodnota ve vzdělávání	6
I.2 Modely ke zjišťování přidané hodnoty	7
I.3 Modely přidané hodnoty učitelů	8
I.3.1 Počátky vývoje modelů přidané hodnoty učitelů	8
I.3.2 Kontroverze ohledně měření přidané hodnoty učitelů	10
I.4 Modely přidané hodnoty jako jeden z typů evaluačních modelů	11
I.5 Metodologické otázky spojené s odhadováním přidané hodnoty učitele	13
I.5.1 Relevance testového nástroje	13
I.5.2 Technické parametry testového nástroje	13
I.5.3 Přřazování učitelů žákům	14
I.5.4 Problematika neměřených proměnných	15
I.5.5 Chyby měření	15
I.5.6 Stabilita modelových odhadů	17
I.5.7 Velikost účinku přidané hodnoty učitele a délka jeho působení na žáky	19
I.6 Jak souvisí přidaná hodnota učitele s jeho dalšími charakteristikami	20
I.6.1 Souvislost mezi přidanou hodnotu učitele a jeho osobními charakteristikami	20
I.6.2 Souvislost mezi přidanou hodnotu učitele a jinými indikátory jeho výkonu	21
I.7 Rozšíření klasických modelů	22
I.7.1 Modely měřící přidanou hodnotu učitelů na středních školách	23
I.7.2 Modely	23
I.7.3 Modely měřící přidanou hodnotu učitelů v jiných zemích	24
I.7.4 Modely měřící přidanou hodnotu v neučitelských profesích	25
I.8 Využití přidané hodnoty učitelů v politikách	26
I.9 Literatura	27
2. Modelování přidané hodnoty – CFR-I a CFR-II	32
2.1 Míra zkreslení odhadů přidané hodnoty učitele	32
2.1.1 Data, metodologie, shrnutí výsledků	33
2.1.2 Diskuze výsledků v kontextu dosavadní odborné literatury	37
2.1.3 Závěr	38
2.2 Zlepšují učitelé s vysokou přidanou hodnotou dlouhodobé výsledky žáků	39
2.2.1 Data, metodologie, shrnutí výsledků	39
2.2.2 Relevance výsledků pro vzdělávací politiku	43
2.2.3 Závěr	45

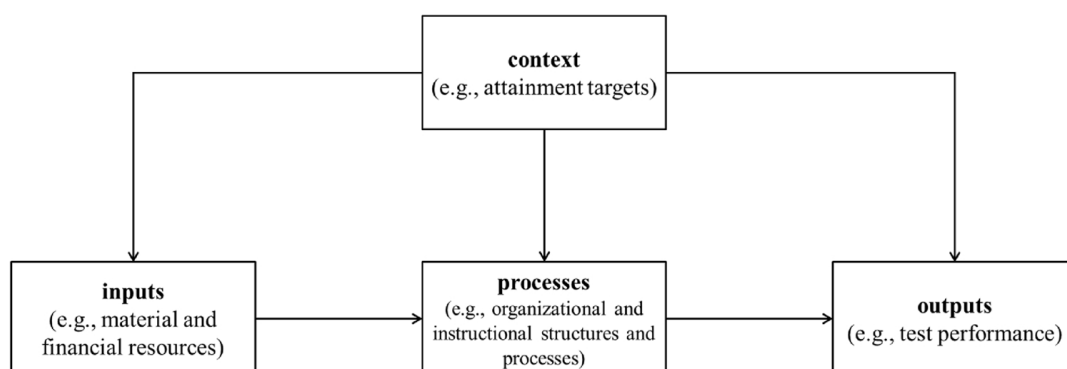
2.3 Replikace CFR-I a CFR-II měření	46
2.4 Literatura	47
3. Experimentální porovnání tří metod hodnocení výkonu učitelů	50
3.1 Dosavadní studie zaměřené na prediktivní validitu metod hodnocení práce učitelů	51
3.2 Design výzkumu	53
3.2.1 Měření	53
3.2.2 Popis experimentu s náhodným přiřazením učitelů a žáků	56
3.3 Empirická strategie	57
3.3.1 Predikce očekávání výkonu učitele	57
3.3.2 Porovnání očekávaného výkonu se skutečnými výsledky po náhodném přiřazení	57
3.4 Shrnutí hlavních empirických zjištění	58
3.5 Důsledky pro formování politik a závěr	59
3.6 Literatura	61
4. Hodnocení výkonu učitelů prostřednictvím jejich přidané hodnoty v praxi škol	64
4.1 Rozdíly mezi učiteli	65
4.2 Cíle studie	65
4.3 Udržení a výkon učitele	66
4.4 Fluktuace učitelů a výsledky žáků	70
4.5 Možnosti implikace	73
4.6 Literatura	73

I. Zjišťování přidané hodnoty ve vzdělávání^{1, 2}

I.1 Přidaná hodnota ve vzdělávání

Termín „přidaná hodnota“ (*value-added*) si výzkumníci v oblasti vzdělávání vypůjčili z ekonomie, kde se používá v souvislosti s částkou, o kterou se v každé fázi výrobního procesu zvyšuje hodnota předmětu (Bacher-Hicks, Koedel, 2023; Koedel a kol., 2015). Ve vzdělávání se tento termín používá v přeneseném smyslu, např. v souvislosti s identifikací lidského kapitálu, který se jedinci podařilo naakumulovat k určitému okamžiku (např. na konci období $t-1$), a poté k odhadu nárůstu tohoto kapitálu prostřednictvím působení různých vzdělávacích vstupů (*inputs*) v období t (ibid.). Přitom za vstupy bývají považovány všechny prostředky, které jsou ve vzdělávacím systému používány k dosahování vzdělávacích cílů. Jedná se například o počty učitelů, školských zařízení, množství učebních materiálů včetně finančních prostředků používaných ke vzdělávání (ibid.).

Obrázek 1: Vstupy, procesy, výstupy a kontextový rámec



Poznámka: Obrázek je upraven dle Scheerense a kol. (2003, str. 18) a převzat z Hofer a kol. (2020).

Obrázek 1 znázorňuje způsob, jakým Scheerense a kol. (2003, citováno z Hofera a kol., 2020) vnímají transformaci vzdělávacích vstupů (*educational inputs*) na vzdělávací výstupy (*educational outputs*), za které jsou pova-

¹ Tento text je podrobným výtahem z níže uvedených textů, ale čerpá i z dalších uvedených zdrojů. Z uvedených textů a zdrojů je přebírán i odkazový aparát:

- Bacher-Hicks, A., Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. Pp. 93–134 in Hanushek, Eric A., Machin, S., Woessmann, L. (eds.) *Handbook of the Economics of Education. Volume 6*. San Diego: Elsevier Science & Technology.
- National Research Council. (2010). National Academies of Sciences, Engineering, and Medicine. *Getting Value Out of Value-Added: Report of a Workshop*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/12820>, a
- Hanushek, E. A., Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American economic review*, 100(2), 267–271.

² Text neprošel jazykovou korekturou.

žovány zejména výsledky žáků ve vzdělávání, jejich znalosti a dovednosti, ale také to, jak je tento proces ovlivňován širším kontextem a procesy na úrovni školy a/nebo třídy (Hofer a kol., 2020).

1.2 Modely ke zjišťování přidané hodnoty

K identifikaci síly vlivu různých vzdělávacích vstupů, které v určitém časovém úseku působí na žáka, se používají statistické modely přidané hodnoty, nebo také statistické modely ke zjišťování přidané hodnoty (*value-added models – VAMs*)³. Jedná se o složité modely, jejichž prostřednictvím lze identifikovat, jakou část zlepšení výsledků žáka lze připsat vzdělávacím vstupům, například škole, učiteli nebo vzdělávacímu programu (National Research Council, 2010). Tyto modely umožňují hledat odpovědi na otázky typu: „*Jakým způsobem přispěla určitá škola (nebo učitel) ke zlepšení výsledku žáka v porovnání s průměrnou školou (nebo průměrným učitelem)*“ (ibid., s. 4). K identifikaci přidané hodnoty je však potřeba izolovat účinky školy, učitele nebo vzdělávacího programu od jiných možných vlivů. Z tohoto důvodu se při modelování přidané hodnoty (*value-added modeling*) berou se v potaz opakované měřitelné výsledky žáka, a to alespoň za dva školní roky, společně s dalšími možnými proměnnými na úrovni žáka a školy (např. míra chudoby, rodinné zázemí žáka, kvalita vedení školy). Jak uvádí Ryška (2008, s. 2), základní myšlenkou tohoto modelování je: „*posoudit efektivitu fungování vzdělávacího systému, škol, učitelů a vzdělávání žáků podle toho, o kolik se zvýší během sledované doby měřitelné výstupy vzdělávání, jako jsou znalosti či dovednosti. Přitom jde o to, abychom co nejlépe postihli, jak tomuto nárůstu přispěje škola či jednotliví učitelé. To zároveň znamená dokázat co nejlépe očistit skutečný příspěvek školy, resp. učitelů od jiných faktorů, které mají vliv na žákovu učení*“. Na základě rešerše literatury včetně zdrojů OECD Ryška (2008, s. 3) předkládá dvě klíčové definice modelů přidané hodnoty:

- „*Modely přidané hodnoty měří příspěvek školy k pokroku žáka vůči předem určeným školním vzdělávacím cílům. Příspěvek je očištěná hodnota od jiných faktorů, které také přispívají k pokroku žáka v učení.*“
- „*Modely zjišťování přidané hodnoty jsou třídou statistických modelů, které se užívají k určení odhadu příspěvku školy k žákově učení měřenému pomocí trajektorií testových skóre.*“

Jinými slovy, modely přidané hodnoty slouží k odhadu přidané hodnoty různých vzdělávacích vstupů (učitelů, školy, vzdělávacích programů), které ve sledovaném období, zpravidla mezi dvěma znalostními standardizovanými testy, na žáky působily. V praxi se používají různé statistické postupy, aby se z předchozích výsledků žáka ve znalostních testech předpověděly jeho budoucí výsledky. Přitom jsou v odhadovaných skóre každého žáka zpravidla zohledňovány různé proměnné (např. předchozí výkon žáka, jeho socioekonomický status, rasa/etnicita, příslušnost), úroveň učitele (např. certifikace, roky praxe, nejvyšší získaný titul, vyučovací postupy, výukové materiály, osnovy) a úroveň školy (např. velikost, typ, nastavení). To, které proměnné jsou nakonec do modelu zahrnuty, závisí na konkrétním typu použitého modelu⁴. Skutečné skóre žáka v testu je následně porovnáno s odhadovaným skóre. Předpokládá se, že rozdíl mezi odhadovaným a skutečným skóre, pokud existuje, je způ-

³ Alternativně lze označovat jako: *value-added measurement, value-added analysis and value-added assessment*.

⁴ Podrobněji viz: https://en.wikipedia.org/wiki/Value-added_modeling#cite_note-25.

soben různými vzdělávacími vstupy školy a/nebo učitele, protože modely jsou konstruovány tak, aby izolovaly přidanou hodnotu školy/učitele/programu od faktorů mimo školu, o nichž je známo, že silně ovlivňují výkon žáka v testu (např. obecná inteligence žáka, míra jeho chudoby, angažovanost rodičů ve vzdělávání). Agregováním výsledků žáků lze následně určit, o kolik zlepšil výsledky žáků konkrétní učitel v porovnání s tím, o kolik by je zlepšil průměrný či jiný učitel.

Statisticky jsou modely přidané hodnoty lineárními regresními modely, ve kterých se odhaduje vliv přidané hodnoty na výsledky žáka v čase t při kontrole vlivu dalších faktorů, kterými jsou výsledky žáka v testu v čase $t-1$ (případně i výsledky starších testů), schopnosti/nadání žáka, a další faktory na individuální úrovni žáka, rodiny a školy. V různých studiích se používají odlišné kontrolní proměnné. Na individuální úrovni a úrovni rodiny jsou běžně používanými proměnnými rasa/etnicita, pohlaví, chudoba (měřena prostřednictvím statusu žáka, který má nárok na oběd zdarma či za sníženou cenu (*free / reduce-price lunch status*), jazyk (žák s odlišným mateřským jazykem), status žáka se zvláštními vzdělávacími potřebami, či vzdělání rodičů. Proměnné na úrovni škol nebo tříd pak bývají zpravidla vypočítány jako agregované ukazatele z údajů o žácích školy/třídy. V některých modelech je vliv některých proměnných měřen jako *fixed-effect*, tedy jako efekt proměnných, které se v průběhu času buď nemění (např. mateřský jazyk, pohlaví, rasa), nebo se mění v čase konstantním způsobem (např. věk). Za takto neměnnou proměnnou se zpravidla také považují nadání žáka nebo kvalita školy (ta se sice může měnit, ale zpravidla ne během krátkého období, v němž probíhá opakované testování jednotlivých žáků).

1.3 Modely přidané hodnoty učitelů

Modely přidané hodnoty slouží k odhadu vlivu přidané hodnoty různých vzdělávacích vstupů na výsledky žáka. V praxi jsou ovšem nejčastěji používány k odhadu vlivu **přidané hodnoty učitele** na výsledky jeho žáků. Tato praxe je motivována řadou výzkumů, které konzistentně dokumentují, že se kvalita učitelů, měřená jejich přidanou hodnotou, významně liší, a to jak mezi školami, tak i v rámci škol. Zároveň řada výzkumů zároveň naznačuje, že standardně používané ukazatele kvality učitelů (např. dosažená kvalifikace, kvalita vystudované školy), které se dlouhodobě používají jako vstupní informace k přijímání do profese, či k určování platu, tuto variabilitu v kvalitě učitelů vysvětlují jen velmi málo. To naznačuje, že dříve sledované charakteristiky učitelů dostatečně nereprezentují jejich kvalitu (Hanushek, Rivkin, 2010).

1.3.1 Počátky vývoje modelů přidané hodnoty učitelů

Předpokládá se, že s myšlenkou posuzovat efektivitu učitele na základě studijních výsledků jeho žáků přišel jako první v roce 1971 Eric Hanushek⁵ (Green, 2014; McCaffrey, Lockwood, 2008), který se zajímal o efektivitu

⁵ Eric A. Hanushek je vedoucím pracovníkem na Hoover Institution of Stanford University. Je lídrem ve vývoji ekonomických analýz vzdělávacích otázek a jeho práce v oblasti efektivit, využívání zdrojů a ekonomických výsledků škol často vstupovaly do tvorby národní i mezinárodní vzdělávací politiky. Na problematice přidané hodnoty učitelů Hanushek často spolupracoval se S. G. Rivkinem – profesorem ekonomie a vedoucím katedry ekonomie na Amherst College, zástupcem ředitele projektu Texas Schools Project na Texaské univerzitě v Dallasu a členem výzkumného týmu CALDER (převzato z Hanushek, Rivkin, 2010).

fungování vzdělávacích systémů. Hanusheka následovala řada dalších autorů, viz například Richard Murnane (1974)⁶, který zkoumal vztah mezi vstupy školy a kognitivními výkony žáků a došel k závěru, že zásadní vliv na výsledky žáků mají jejich učitelé.

Významný vliv na rozvoj konceptu přidané hodnoty učitelů a jeho zavedení do praxe měl William Sanders⁷. Sanders v 90. letech 20. století vyvinul model přidané hodnoty učitelů (TVVAS/EVAAS)⁸, který měl sloužit jako nástroj pro hodnocení efektivity učitelů. Následně tento nástroj v Severní Karolině a Tennessee implementoval do praxe. Jeho cílem bylo prokázat, že zjišťování přidané hodnoty učitelů je účinnou a zároveň efektivní metodu pro určování vlivu jednotlivých učitelů na míru akademického růstu jejich žáků (Sanders, Rivers, 1996). Sanders předpokládal, že tento nástroj bude možné při dodržení určitých podmínek využít k odlišení vysoce efektivních učitelů od učitelů průměrných a neefektivních. Zájem o tuto problematiku následně významně vzrostl po zveřejnění technické zprávy (Sanders, River, 1996), která ukázala, že efektivita učitele odhadovaná z výsledků předchozích testů jeho žáků předpovídá výsledky těchto žáků nejméně dva roky dopředu. Toto zjištění naznačuje, že vliv učitelů na výsledky žáků může být dlouhodobý a akumulace těchto účinků může hrát podstatnou roli ve vzdělávání (ibid.).

V následující studii se Sanders a jeho kolegové (Sanders, Wright, Horn, 1997) pokusili identifikovat relativní velikost vlivu učitelů na výsledky jejich žáků za současného zvažování vlivu heterogenity uvnitř třídy, úrovně výsledků žáků a velikosti třídy. Výsledky jejich studie naznačují, že učitelé hrají v dosažených výsledcích žáků dominantní úlohu, a že kontextové proměnné jako je velikost třídy a heterogenita žáků ve třídě, mají na akademický úspěch žáků relativně malý vliv. Na tato zjištění navázaly další výzkumy, které docházely k obdobným závěrům, a to že mezi učiteli existují velké rozdíly v jejich přidané hodnotě, že učitelé s vysokou přidanou hodnotou mají na akademický úspěch žáků nejen krátkodobý, ale i dlouhodobý pozitivní vliv, a že své žáky ovlivňují i v pozdějším životě, a to i v takových oblastech jako je výše jejich mezd, pravděpodobnost studia na vysoké škole apod. (Hanushek, Rivkin, 2010; Chetty a kol., 2014a).

Používání modelu přidané hodnoty se v rámci Spojených států postupně šířilo do dalších států a školských distriktů. K tomu pomohlo schválení zákona *No Child Left Behind* v roce 2001 (Braun, 2005). Ve snaze zlepšit výsledky žáků byla tímto zákonem zavedená povinnost jejich průběžného monitorování prostřednictvím pravidelného testování žáků 3. – 8. ročníků základních škol a prvních ročníků středních škol. Tímto krokem se zlepšila dostupnost dat o výsledcích žáků a otevřel se prostor pro modelování přidané hodnoty učitelů (National Research Council, 2010).

Modelování přidané hodnoty v posledních desetiletích přitahuje značnou pozornost, a to především těch, kteří se zajímají o problematiku zlepšování škol, výuky, vzdělávacích programů, stejně tak jako problematiku zodpovědnosti škol a učitelů za výsledky žáků a v důsledku toho i problematikou měření efektivity škol a učitelů (National Research Council, 2010). Modelování přidané hodnoty je dnes považováno za jeden z klíčových nástrojů výzkumníků, jejichž cílem je pochopit, které faktory ovlivňují efektivitu vzdělávacích systémů a vzdělá-

6 Působil na Harvardské univerzitě.

7 Americký statistik a výzkumný pracovník z University of North Carolina at Chapel Hill.

8 Tzv. Tennessee Value-Added Assessment System (TVAAAS), také známý jako Educational Value-Added Assessment System (EVAAS).

vání (Koedel a kol., 2015). Vychází se přitom z presumpce, že pro budování efektivních vzdělávacích systémů je klíčová znalost toho, jak vzdělávací vstupy ovlivňují vzdělávací výstupy (Tierney a kol., 2022). Především pak povědomí o tom, jak k růstu znalostí a dovedností žáků přispívá škola a učitelé v porovnání s jinými známými faktory ovlivňujícími učení žáků, které ovšem škola nemůže ovlivnit (Ryška, 2008).

1.3.2 Kontroverze ohledně měření přidané hodnoty učitelů

Měření přidané hodnoty učitelů ovšem v praxi doprovázejí kontroverze. Ty vzbuzuje zejména jejich aplikace ve veřejné politice a jejich používání k hodnocení učitelů (Koedel a kol., 2015) a následným personálním rozhodnutím. Používání přidané hodnoty k tomuto účelu totiž vyvolává řadu obav, z nichž Koedel a kol. (2015) zmiňují dvě nejvýznamnější:

(1) odhady přidané hodnoty učitelů mohou být zkreslené (Baker a kol., 2010, Paufler, Amrein-Beardsley, 2014, Rothstein, 2009, 2010) a

(2) odhady přidané hodnoty se zdají být příliš nestabilní na to, aby mohly být použity pro vysoce důležitá personální rozhodnutí jako je např. zaměstnávání učitelů apod. (Baker a kol., 2010, Newton a kol., 2010).

Kritici také tvrdí, že používání modelů přidané hodnoty k hodnocení jednotlivých učitelů nebylo vědecky dostatečně ověřeno. Předpokládají, že výsledky mohou být ovlivněny náhodou nebo podmínkami, které učitel nemůže ovlivnit, jako je například externí doučování žáka (Dillon, 2010). Určité pochybnosti o tom, zda je vhodné učitele hodnotit na základě jejich přidané hodnoty, která je odvozená z výsledků standardizovaných testů jejich žáků, vyjádřil i expertní tým pracující pro *Economic Policy Institution*, který v roce 2010 vydal k této problematice výzkumnou zprávu (Baker a kol., 2010), ve které doporučuje více komplexní hodnocení učitelů, ne spoléhání se na jeden ukazatel, protože neexistuje jeden dokonalý způsob hodnocení učitelů. Autoři studie se ohrazují především proti tomu, aby byla na základě výsledků žáků činěna důležitá personální rozhodnutí. Domnívají se, že hodnocení učitelů by mělo být mnohem komplexnější. Jejich základem by mělo být hodnocení učitele kompetentními supervizory a kolegy, hodnocení založená na měření studijních výsledků žáků by měla hrát pouze doplňkovou roli.

Výhrady k používání modelů s přidanou hodnotou k hodnocení učitelů vyjadřuje i Bracey (2013). Problém vidí v tom, že je možné modely přidané hodnoty považovat za kruhové modely, protože definují efektivní učitele jako ty, kteří zvyšují skóre v testech, a následně používají vyšší skóre v testech k určení efektivnějších učitelů. Používání modelů s přidanou hodnotou v oblasti vzdělávání zkritizovala i *American Statistical Association (ASA)*, která v roce 2014 vydala prohlášení, ve kterém kritizovala omezenost vstupních dat, vliv faktorů, které nebyly do modelů zahrnuty a velikost standardní chyby, což vše může vést k nestabilitě meziročních hodnocení. Na otázky ASA následně v odborné literatuře reagovali Chetty a kol. (2014b), kteří dokumentovali, které z otázek ASA již byly vyřešeny, a které by si zasloužily další výzkumnou pozornost. S kritikou modelů přidané hodnoty přišel i Ewing (2011). Obdobně jako ASA kritizoval vstupní data a možný vliv faktorů, které nebyly do modelu zahrnuty. *National Research Council* (2010) navíc upozorňuje, že v některých případech může docházet k tomu, že na základě

analýzy přidané hodnoty jsou některé školy, učitelé nebo programy označené jako vysoce efektivní – „nejlepší“, ale nemusí být „nejlepší“ při použití jiných kritérií, protože model přidané hodnoty dává větší váhu určitému skóre testu než jiným možným sledovaným charakteristikám efektivity (National Research Council, 2010).

Další skupina problémů s používáním modelů přidané hodnoty k hodnocení efektivity učitelů na základě výsledků jejich žáků spočívá v tom, že zlepšování výsledků žáků lze definovat mnoha způsoby, například prostřednictvím jejich výsledků ve standardizovaných testech, změnou podílu žáků, kteří splnili, nebo překročili stanovené standardy, nebo rozdílem mezi skutečnými a očekávanými průměrnými výsledky žáků. Volba způsobu, jakým bude sledováno zlepšování výsledků žáků, je přitom rozhodující pro dosažení požadovaných dopadů a každá volba má svá omezení a je spojená s určitými kompromisy. Pokud bude například za kritérium efektivity učitele zvoleno zlepšování průměrného skóre žáků v testech, učitel, který zlepší výsledky svých žáků více než ostatní učitelé s podobnými žáky, bude považován za efektivnějšího, přestože žáci nemusí v testech dosahovat požadované úrovně výsledků. Je tedy potřeba si uvědomit, jaké limity mají různé způsoby definování efektivity učitele, a to ještě koncipováním různých systémů jejich hodnocení, protože tím se zvýší pravděpodobnost, že bude celý systém hodnocení efektivity učitelů navržen koherentně a bude mít vyšší šanci na dosažení požadovaných cílů. Rovněž je potřeba vzít v potaz, že v praxi jsou cíle tvůrců politik obvykle ambicióznější, než je možné s existující statistickou metodologií a kvalitou dat dosáhnout (National Research Council, 2010).

1.4 Modely přidané hodnoty jako jeden z typů evaluačních modelů⁹

Modely přidané hodnoty ovšem nepřinášejí důležité informace pouze o přidané hodnotě učitelů, ale také o kvalitě a zlepšování škol, stejně tak jako o efektivitě a kvalitě celého vzdělávacího systému. V důsledku toho bývají považovány i za jeden z typů **evaluačních modelů** (National Research Council, 2010). National Research Council (2010) upozorňuje, že ve vzdělávání se používá několik typů evaluačních modelů, které jsou obdobně jako modely přidané hodnoty založeny na výsledcích žáků ve znalostních testech, a to zejména modely aktuálního stavu (*status model*), modely změn mezi jednotlivými kohortami žáků (*cohort-to-cohort change models*), modely sledování růstu / zlepšování výkonu žáků (*growth models*).

Ad1) Modely aktuálního stavu poskytují přehled o výkonech žáků v určitém okamžiku. Výkony mohou být následně srovnány se stanovenými vzdělávacími cíli. Tyto modely jsou užitečné zejména v případě, že jsou hledány odpovědi na otázky typu, jaké procento žáků (státu, školního distriktu, školy, učitele, určité specifické skupiny žáků) dosáhlo v daném školním roce požadovaných cílů. Významné jsou také v případě, že se snažíme identifikovat výkon školy v komparaci se stanovenými cíli vzdělávání (National Research Council, 2010).

Ad2) Modely změn mezi jednotlivými kohortami jsou pro jednotlivé státy, školní distrikty, školy a učitele důležitým zdrojem informací o změnách ve výkonech žáků ve dvou různých okamžicích. Nejedná se ovšem

⁹ Tato kapitola je podrobným výtahem z kapitoly z textu: National Research Council. (2010). National Academies of Sciences, Engineering, and Medicine. 2010. Getting Value Out of Value-Added: Report of a Workshop. Washington, DC: The National Academies Press. <https://doi.org/10.17226/12820>. Z tohoto textu je přebíráno i odkazový aparát.

o výkony stejných žáků, ale o výkony žáků určitých ročníků. Například lze porovnat procento žáků 4. třídy, kteří ve čtení ve sledovaném školním roce dosáhli požadovaných výkonů, s procentem žáků, kteří stejných výkonů dosáhli ve 4. třídě v předchozím školním roce. Tyto modely nám poskytují odpovědi na otázku, zda jsou výkony žáků sledovaného ročníku v daném školním roce lepší než výkony žáků stejného ročníku v předchozích letech.

Ad3) Modely sledování růstu / zlepšování výkonu se používají ke sledování změn ve výkonech stejných žáků v čase, z jednoho roku na druhý. Cílem těchto modelů je určit rozsah pokroku žáků v čase. Přitom rozsah pokroku (např. současných žáků 4. ročníku) lze následně porovnat s výsledky stejné skupiny žáků v loňském roce, tedy v době, kdy navštěvovali 3. ročník. Tento typ modelu je vhodný, pokud chceme zjistit, jak moc se v průměru zlepšil výkon žáků mezi jednotlivými ročníky. Používáním těchto modelů lze např. ocenit výkon škol/učitelů, jejichž žáci vykazují zlepšení výkonu, a to především v případě, že byl jejich výkon v předchozích letech malý. Určitým limitem těchto modelů je ovšem skutečnost, že obvykle nekontrolují faktory z prostředí žáka nebo školy. Tedy neumí explicitně identifikovat, které faktory jsou ve skutečnosti zodpovědné za zlepšování výkonu žáků.

Ad4) Modely přidané hodnoty jsou složité statistické modely, které se pokoušejí připsat určitou část zlepšování výsledků žáků v průběhu času školám, učitelům nebo vzdělávacím programům. Tyto modely hledají odpovědi na otázku, jak školy/učitelé/programy přispívají ke zlepšení výsledků žáků v porovnání s průměrnou školou/učitelem/programem. Nebo ekvivalentně, jak velkou část změny ve výkonech žáků lze připsat škole/učiteli. Tyto modely zároveň extrahují účinky školy/učitele/vzdělávacích programů od dalších možných proměnných na úrovni žáka či školy. Za odhad přidané hodnoty školy nebo učitele je pak považován rozdíl mezi pozorovaným zlepšením žáků a očekávaným zlepšením. I když u některých modelů nemusí být tento vztah tak přímočarý, ve všech případech odhad přidané hodnoty slouží k identifikaci příspěvku školy, učitele, vzdělávacího programu k výkonu žáků (National Research Council, 2010)¹⁰.

Přitom použití každého z výše uvedených typů modelů má svoje výhody a omezení. Například v případě, že se tvůrci evaluačních systémů zajímají především o to, aby všichni žáci dosáhli určité úrovně znalostí a dovedností, pak mohou být za vhodné považovány modely aktuálního stavu, jejichž používání nařizuje legislativa *No Child Left Behind*. Používání těchto modelů má však svoje limity, a to především v případě, že by měly sloužit k hodnocení škol. Mohlo by docházet k situacím, kdy školy, které ke studiu přijaly žáky s dobrými výsledky ve standardizovaných znalostních testech, ale mají nízké skóre přidané hodnoty, by byly pozitivně hodnoceny (nebo alespoň ne-sankcionovány), zatímco školy, které přijaly žáky s nízkými výkony v těchto testech, ale mají vysoké skóre přidané hodnoty, by mohly být identifikovány jako školy vyžadující zlepšení, případně dokonce sankcionovány (National Research Council 2010). Ve skutečnosti by se ovšem jednalo o efektivní školy, které významně zlepšují výsledky svých žáků (ibid.).

¹⁰ National Research Council (2010) upozorňuje, že existuje další kategorie modelů, která je velmi podobná modelům přidané hodnoty, ale na rozdíl od těchto modelů nepracuje s výsledky předchozích testů žáků. Tyto modely se označují jako upravené modely aktuálního stavu (*adjusted status models*) a používají statistické techniky k „úpravě“ průměrného prospěchu žáků napříč jednotkami analýzy (tj. školami, učiteli nebo programy) zohledněním rozdílů ve složení žáků a jiných faktorů. Ve skutečnosti tyto modely porovnávají výsledky žáků pro podobné jednotky. Pokud například žáci, jejichž rodiče mají vysokoškolské vzdělání, mívají vyšší skóre v testech než žáci, jejichž rodiče mají nižší dosažené vzdělání, pak průměrné výsledky žáků školy s vyšším procentem vysokoškolsky vzdělaných rodičů budou upraveny směrem dolů, zatímco průměrné skóre žáků škol s nižším procentem vysokoškolsky vzdělaných rodičů bude upraveno směrem nahoru. Jak upozorňuje National Research Council (2010), takové modely jsou prvním krokem ke skutečným modelům s přidanou hodnotou, ale nevyužívají cenné informace o předchozím výkonu žáků.

National Research Council (2010) doporučuje před zavedením evaluačního systému do praxe zvážit výhody a limity všech evaluačních modelů a tím zvýšit pravděpodobnost, že bude celý systém navržen koherentně a bude mít větší šanci na úspěch a dosažení požadovaných cílů. Zároveň ovšem upozorňují, že „v praxi jsou cíle tvůrců politik obvykle ambicióznější, než může statistická metodologie a kvalita dat podpořit“ (s. 6).

1.5 Metodologické otázky spojené s odhadováním přidané hodnoty učitele

Schopnost modelů spolehlivě odhadnout přidanou hodnotu učitelů samozřejmě závisí na řadě faktorů a jejich používání je spojeno s potřebou řešit řadu metodologických otázek a problémů. V odborné literatuře je proto možné najít rostoucí počty prací, které se těmito otázkami zabývají (Jacob, Lefgren, Sims, 2008; Kane, Staiger, 2008; Ishii, Rivkin, 2009; Rothstein, 2010; Hanushek, Rivkin, 2010; National Research Council and National Academy of Education 2010; Koedel, Rockoff, 2015; Bacher-Hicks, Koedel, 2023). Metodologické otázky, nad kterými se autoři studií zamýšlejí, mají nejrůznější charakter. Zahrnují otázky velmi obecné i otázky úzce zaměřené na řešení specifických problémů jednotlivých metod. V následující části textu bude podán přehled o základních metodologických otázkách, které je při měření přidané hodnoty učitele třeba zvážit.

1.5.1 Relevance testového nástroje

Jednou ze základních obecných metodologických otázek spojených s měřením přidané hodnoty učitelů je skutečnost, že klíčovým zdrojem informací o efektivitě učitelů jsou výsledky jejich žáků ve standardizovaných testech znalostí a dovedností. První otázkou tedy je, zda tyto testy skutečně měří to, co učitelé ve svých hodinách vyučují. Pokud by testy v dostatečné míře neměřily to, co učitelé skutečně učí, jakkoliv sofistikovaná metoda, která by jako vstupní data používala výsledky těchto testů, by z logiky věci mohla jen omezeně vypovídat o vlivu, jaký mají na výsledky žáků v testech jejich učitelé (National Research Council, 2010).

Důležitá je také otázka, zda tyto testy skutečně měří znalosti a dovednosti, které lze považovat za důležité nebo cenné. Navíc je potřeba si uvědomit, že testování zpravidla probíhá pouze v některých předmětech/oblastech – nejčastěji v matematice, ve čtenářských dovednostech či v přírodních vědách – zatímco jen zřídka kdy se testují znalosti a dovednosti v sociálních vědách, umění nebo IT. Testy také neměřící další důležité kvality, které se školy snaží předat žákům, jako jsou intelektuální zvědavost, motivace, vytrvalost při zvládání obtížných úkolů a překonávání překážek, nebo schopnost spolupracovat s ostatními (National Research Council, 2010).

1.5.2 Technické parametry testového nástroje

Existují ovšem další otázky, které musí být řešeny, a to především jakým způsobem je úroveň znalostí a dovedností měřena. Například zda existuje nějaká maximální možná horní hranice možného výkonu žáka v testu, či intervalová škála pro výsledky žáka v testu, nebo zda je zlepšení žáka v testu vyjádřené jeho posunem

v bodovém skóre, například z 200 na 300 u škály v rozsahu 0–800 bodů (nebo z percentilu 5 do percentilu 15) stejně hodnotné jako zlepšení žáka z bodového skóre 600 na 700 bodů na téže škále (nebo z percentilu 80 do percentilu 90), víme-li, že výsledky testů ve skutečnosti nemají vlastnosti intervalové škály (National Research Council, 2010)?

Studie Pophama (2007) rovněž upozornila na to, že různé testy, jejichž výsledky jsou používány jako vstupní data pro odhady přidané hodnoty učitele, se významně liší v tom, nakolik jsou citlivé na správně podané instrukce před zahájením testu. Výsledky některých testů totiž příliš nezávisí na tom, jak kvalitní instrukce byly žákům poskytnuté při zadávání testů, zatímco jiné ano. Použití testů citlivých na kvalitu instrukcí poskytovaných při zadávání testu vnáší do výsledků vliv další proměnné, která nesouvisí se samotnou přidanou hodnotou učitele.

Tyto a další podobné technické specifikace používaných testů jsou ve svém důsledku pro měření přidané hodnoty podstatné, protože odhadovaná přidaná hodnota závisí i na technických detailech zvoleného testu. Rozdíly ve výsledcích získaných za pomoci různých technik testů znalostí a dovedností existují, ale nejsou zase natolik významné, aby zásadně ovlivnily hlavní výsledky (Hanushek, Rivkin, 2010). Ačkoliv výsledky výzkumů naznačují, že si tyto záležitosti zaslouží pozornost, nezdá se, že by ohrožovaly základní výsledek studií, což je zjištění, že kvalita učitelů se podstatně liší. Různé technické detaily používaných testů znalostí a dovedností žáků vedou ke stejnému základnímu zjištění a možná odlišit učitele s největší přidanou hodnotou od ostatních učitelů.

1.5.3 Přiřazování učitelů žákům

Ideální metoda modelování přidané hodnoty učitelů je založena na principu, že žáci mohou být při použití randomizovaného experimentálního designu měření učitelům přiřazováni náhodně, aby bylo možno měřit efekt, jaký mají konkrétní učitelé na jimi vyučované žáky a jak se výsledky žáků v testech mění v závislosti na tom, k jakému učiteli byli žáci přiřazeni. Při interpretaci výsledků je ovšem nutno brát v úvahu, že v reálném životě nikdy nejsou učitelé a žáci k sobě přiřazováni náhodně. To je zvláště patrné na úrovni jednotlivých škol, kde je přiřazení žáků do konkrétní školy v mnoha zemích zásadně ovlivněno její místem bydliště a vymezením jejího spádového obvodu, což do procedury přiřazování učitelů a žáků vnáší významný vliv sociálně-prostorových nerovností v dané zemi, i v jednotlivých regionech a městech. Studie navíc upozorňují, že je potřeba brát v úvahu i skutečnost, že přiřazování učitelů k jednotlivým žákům neprobíhá náhodně dokonce ani v rámci jedné školy (Bacher-Hicks, Koedel, 2023). „Dobří žáci“, tj. v logice měření ti, kteří dosahují dobrých výsledků ve standardizovaných testech, se jimi mohou stát proto, že měli dobré učitele (identifikace takových učitelů je cílem měření přidané hodnoty učitele), ale nelze vyloučit, že skutečnost, že jsou dobří žáci nad reprezentování ve třídách dobrých učitelů, může být částečně způsobeno i jejich vzájemným nenáhodným přiřazováním (dobří žáci se snaží vybírat si dobré učitele, nebo si dobří učitelé snaží vybírat dobré žáky, případně k takovému přiřazování dochází vlivem působení nějakého jiného faktoru).

1.5.4 Problematika neměřených proměnných

Specifickou metodologickou otázkou, která souvisí s odhadem přidané hodnoty učitelů, je, nakolik může vést vynechání některé z teoreticky vlivných proměnných z modelu, ke zkresleným odhadům přidané hodnoty učitelů. Používaná metodologie pro měření přidané hodnoty učitelů se vždy snaží odlišit „čistý vliv“ pedagogické práce učitele/školy na výsledky žáků ve standardizovaných testech od vlivů ostatních možných faktorů, které působí na úspěšnost žáků v testech. Je zřejmé, že do modelů odhadujících samotnou přidanou hodnotu není možné zahrnout vliv všech existujících faktorů. U některých z nich je to proto, že jejich vliv se dá objektivně předpokládat, ale jejich efekt se nedá snadno, případně vůbec, změřit. U jiných faktorů předpokládáme, případně z provedených výzkumů víme, že jejich vliv je malý, a proto není z hlediska porovnání nákladů a zisků smysluplné jejich vliv měřit a do modelů je zahrnovat. Nelze vyloučit ani možnost, že vliv některých faktorů může být potenciálně významný, ale v modelech nejsou brány v potaz jednoduše proto, že vliv těchto faktorů nebyl dosud identifikován. Konečně vliv některých faktorů je známý, existují způsoby, jak vliv měřit, ale stejně nejsou do modelů zahrnuty, protože existují praktická (například finanční nebo časová) omezení, která brání změření vlivu všech známých faktorů a jejich zahrnutí do modelů. V takovém případě je metodologickým problémem určit, které proměnné neměřit a do modelu nezahrnovat, a spolehlivě odhadnout, jaký vliv takové neměřené proměnné budou mít na celkové výsledky měření. Z odborné literatury je známo, že zvláště pokud empirický model nezohledňuje rozdíly mezi žáky, které ovlivňují jejich výběr školy, mohou být odhady vlivů učitelů a jejich agregované rozptyly zkreslené. Jedná se o obzvláště složité otázky, vezmeme-li v úvahu, že jak rodiče, tak zaměstnanci školy mají možnost volby (viz Hanushek, Kain, Rivkin, 2004). Tyto problémy byly předmětem zájmu výzkumníků již dlouhou dobu a byla o nich napsána řada studií, viz např. Hanushek (1992).

1.5.5 Chyby měření

K měření přidané hodnoty se používají metody, které na základě získaných dat měří hodnotu proměnné charakterizující hodnotu pedagogické práce. Jako u každého měření je potřeba počítat s tím, že hodnoty proměnných jsou měřeny s určitou chybou. Bacher-Hicks a Koedel (2023) se ve své studii věnují problematice systematické chyby měření modelů přidané hodnoty. Připomínají, že v praxi vycházejí modely přidané hodnoty v naprosté většině případů z observačních dat, nejsou to ovšem data pocházející z kontrolovaného randomizovaného experimentu. Aby mohla být observační data použita k odhadu kauzálního vztahu mezi nezávislou (např. vliv učitele) a závislou proměnnou (výsledky žáka ve standardizovaných testech znalostí a dovedností), je třeba, aby byly dodrženy určité předpoklady. Jako ve všech modelech tohoto typu postavených na observačních datech je velikost zkreslení měření (*bias*) závislá na velikosti vlivu nepozorovaných/neměřených faktorů (*unobserved dimensions*).

Zkreslení měření bude zmenšeno či úplně eliminováno, pokud (a) nenáhodné přiřazování učitelů žákům bude limitováno, nebo (b) pokud nenáhodné přiřazování učitelů žákům existuje, ale zároveň jsou k dispozici dostatečně bohaté informace o faktorech, které tuto nenáhodnost přímo ovlivňují, nebo o faktorech vysoce korelovaných s faktory, které nenáhodnost ovlivňují. Empirické studie dokazují, že v praxi nejsou učitelé přiřazováni žákům náhodně. Mezi faktory, které toto přiřazování ovlivňují, můžeme najít i takové, které v typických mode-

lech přidané hodnoty nebývají brány v úvahu (např. interakce mezi žáky a učiteli, preference rodičů). V odborné literatuře existuje několik pokusů, jak testovat vliv běžně neměřených faktorů na zkreslení odhadu u standardních modelů.

Chetty a kol. (2014a) byli schopni do modelů přidat i další proměnné, jako příspěvek zaměstnavatele rodičům žáků na jejich důchod (*401(k) contribution*)¹¹, vlastnictví nemovitosti, příjem domácnosti v kombinaci s rodinným stavem. Následně porovnali rozdíl v odhadu efektu učitele s modelem, který výše uvedené proměnné neobsahoval. Výsledek analýzy prokázal, že přidané proměnné zlepšily přesnost odhadu výsledků testu, což potvrdilo vliv těchto faktorů, rozdíl oproti jednodušším klasickým modelům byl ale relativně malý, jen 0,2 procent naměřené hodnoty.

Jinou možností je udělat odhad zkreslení měření za pomoci modelu užívajícího kvazi-experimentálního designu využívajícího skutečnost, že učitelé fluktuují mezi třídami/školami, takže je možné odhadovat jejich vliv na výsledky žáků v porovnání s vlivem jiných učitelů, kteří je v jejich roli nahradili. Za pomoci tohoto kvazi-experimentálního designu Chetty a kol. (2014a) odvodili, že je zkreslení měření standardních modelů 2,6 procent naměřených hodnot.

Dalším možným způsobem, jak odhadnout velikost zkreslení měření, je provedení experimentální studie. V rámci experimentu je nejprve standardními modely odhadnuta „přidaná hodnota“ konkrétních učitelů, následně jsou žákům náhodně přiděleni učitelé, ti je nějakou relevantní dobu učí, a potom je následně změřen výkon žáků v testech.

Pokud by nebyly odhady přidané hodnoty zatíženy systematickou chybou, měly by být budoucí výsledky testů jejich žáků s akceptovatelnou mírou přesnosti odhadnuty. Experimentální studie (Kane, Staiger, 2008; Bacher-Hicks a kol., 2019) ve všech případech dokázaly, že zkreslení měření není u klasických modelů příliš velké. Neexistuje ovšem odborný konsensus nad tím, zda je vytváření složitějších modelů zahrnujících vliv nenáhodného přidělování učitelů žákům jako další proměnné ku prospěchu věci či nikoli. Empirická analýza Rothsteina (2010) zdůraznila význam nenáhodného přidělování učitelů žákům a výběr tříd žáky. V jeho práci je analyzována možnost, že nenáhodné přiřazení žáků do tříd vede k získání zkreslených odhadů přidané hodnoty učitele. Proti modelům zahrnujícím vliv nenáhodného přidělování ovšem hovoří to, že jsou mnohem složitější než modely, které tento vliv nezahrnují. Pochybnost o jejich užitečnosti ovšem existuje také proto, že bylo empiricky ověřeno, že složitější modely mohou odhadovat klíčové parametry dokonce s menší přesností, než méně komplexní modely (Bacher-Hicks, Koedel, 2023). Bacher-Hicks a Koedel (2023) proto obecně doporučují užití komplexnějších modelů až v situaci, kdy se měří přidaná hodnota učitelů ve velkých školách, v nichž žáky učí hodně různých učitelů a v místech, ve kterých jsou obecně velké rozdíly mezi školami (jak učitelé, tak žáci jsou do různých škol distribuováni značně nenáhodně). Kromě toho existují i metodologická omezení randomizovaných experimentálních studií. Jedním z nich je skutečnost, že ve všech dosud provedených studiích tohoto typu bylo přiřazování učitelů žákům náhodné v rámci školy, ale nebyl

¹¹ Poznámka: plán 401(k) navrhl Kongres Spojených států. Cílem je povzbudit Američany k tomu, aby si spořili na důchod. Jedná se o plán penzijního spoření nabízený mnoha americkými zaměstnavateli, který má pro střadatele různé daňové výhody. Zaměstnanec, který se zaregistruje do 401(k), zároveň vyjadřuje souhlas s tím, že procento z každé výplaty bude vyplaceno přímo na investiční účet. Zaměstnavatel může dorovnat část nebo celý tento příspěvek (Fernando, 2023).

kontrolován vliv nenáhodného přiřazení různých žáků do různých škol. Metody tedy dokázaly měřit jen potenciální zkreslení způsobované v praxi běžně pozorovaným nenáhodným způsobem přiřazování učitelů a žáků v rámci školy, ale nemohly poskytnout informaci o potenciálním zkreslení způsobeném nenáhodným přiřazováním žáků i učitelů do různých škol.

Jiným typem chyby spojené s měřením přidané hodnoty učitele je chyba měření samotného testu, kterým se měří dosažená úroveň znalostí a dovedností žáků, což je komplikace, která nabývá na významu v odhadech rozptylu kvality učitelů založených na residuích (*residual-based estimates*). Žádný znalostní test (*achievement test*) totiž neměří znalosti žáků zcela přesně. Výběr konkrétních otázek, náhodné události doprovázející testovací situace, obeznámenost či neobeznámenost s testy a další faktory mohou totiž vést k tomu, že se naměřené skóre bude lišit od skutečných znalostí žáků, a tato chybovost testů se přenáší do chyb v odhadech přidané hodnoty pro učitele. Další komplikací je to, že chyba měření je různá u různých typů otázek – jiná je u uzavřené otázky, kde žák vybírá jednu z možných odpovědí (tam závisí i na počtu nabízených možností), jiná je u otázek, na které žáci odpovídají volnými odpověďmi, které jsou následně hodnoceny testery. Technickou komplikací je rovněž skutečnost, že chyba měření u celkových výsledků testu není konstantní ve všech částech hodnotící škály. Jednou z charakteristik mnoha testovacích nástrojů je to, že chyba měření je podstatě větší u výsledků na horním a spodním okraji škály, zatímco je mnohem menší ve střední části škály (National Research Council, 2010). Agregátní skóre pro třídu (či školu), ve které je vysoké procento žáků se špatnými výsledky v testech, jsou proto zatížena větší výběrovou chybou, než agregátní skóre v běžných třídách (školách), což může významně ovlivnit i spolehlivost odhadu přidané hodnoty učitele působících v těchto typech tříd (škol).

1.5.6 Stabilita modelových odhadů

Za velmi významnou problematiku modelování s přidanou hodnotou je nutno považovat i stabilitu modelových odhadů (Bacher-Hicks, Koedel, 2023). Bacher-Hicks a Koedel (2023) ji považují za klíčovou charakteristikou modelových odhadů přidané hodnoty učitele. Stabilitou se rozumí konsistence přidané hodnoty učitele v čase. Například stabilita přidané hodnoty učitele matematiky poukazuje na skutečnost, že daný učitel přidává svým žákům stejnou hodnotu k testovým skóre. Pokud je tedy na základě výsledků žáků v testech odhadnuto, že má některý učitel vysokou přidanou hodnotu, patří tedy k „nejlepším učitelům“, bude mít tuto vlastnost i v budoucnu, a i jeho budoucí žáci od něj budou získávat vyšší přidanou hodnotu, než by dostali od průměrného učitele. Bez alespoň elementární míry stability naměřené přidané hodnoty učitele by získané odhady neměly přílišný faktický význam, protože by vypovídaly jen o minulosti a nepřinášely by žádnou informaci o možné budoucnosti. Velký počet empirických studií prokázal, že stabilita přidané hodnoty učitele, měřená jako korelace mezi modelovými odhady z opakovaných měření u jednotlivých učitelů, je sice kladná (tj. učitelé, kteří měli v minulosti naměřenou vyšší hodnotu, mají tendenci mít naměřenou vyšší hodnotu i při opakovaném měření), ale poměrně významně odlišná od čísla 1, pokud je vyjádřena jako Pearsonova korelace (ve studiích se hodnota Pearsonova korelačního koeficientu pohybovala v rozmezí 0,18–0,64). Důvody pro nezanedbatelné změny v naměřené hodnotě přidané hodnoty učitele jsou různé, nejčastěji se udávají tři:

- skutečné změny v efektivitě práce učitele;
- náhodné jedinečné události v daném roce měření;
- chyba měření.

Skutečně naměřená přidaná hodnota práce učitele se může zvyšovat v důsledku zvyšující se efektivity jejich práce například v souvislosti s tím, jak získávají zkušenosti s rostoucí délkou praxe neb dostanou další vzdělávání v rámci profesního rozvoje učitele atd. Jiné změny přidané hodnoty učitele mohou být reálné, ale nemusí to nutně být změny k lepšímu, protože mohou reflektovat proměny pracovního kolektivu učitelů a či vedení školy, které na efektivitu práce učitele mohou působit v pozitivním i v negativním směru.

Příkladem náhodné jedinečné události mající vliv na naměřenou přidanou hodnotu učitele je například přiřazení učitele do kolektivů žáků, kteří se výrazně odlišují od předchozích kolektivů jeho žáků. Jiným důvodem mohou být například náhodné proměny složení testovaných žáků, které nesouvisí s efektivitou práce učitele, ale mají vliv na jejich úspěšnost v testech (například absence nejhorších či nejlepších žáků při testech například z důvodu nemoci).

Je tedy zřejmé, že odhad přidané hodnoty učitele provedený na základě měření výsledků jeho žáků může do jisté míry přirozeně kolísat, protože každé měření je zatíženou jistou náhodnou chybou měření. Velikost náhodné chyby měření je ovlivněna velikostí zkoumaného vzorku žáků/učitelů, se vzrůstající velikostí vzorku se zkreslení zmenšuje. Goldhaber a Hansen (2013) také upozorňují, že zlepšení modelů lze dosáhnout prodloužením časové řady měření, protože zlepšení žáků způsobené přidanou hodnotou učitele je nelineární, a proto ho delší časové řady testování žakovských znalostí a dovedností standardizovanými testy dokážou lépe zachytit. Podle National Research Council (2010) je nejobvyklejším řešením problémů s krátkodobou nestabilitou výsledků měření pracovat nikoliv s údaji z jednoho roku, ale s tříletými klouzavými průměry.

Stabilita přidané hodnoty učitele je také ovlivněna samotným typem modelu použitého k jejímu odhadu a přesnou specifikací konkrétního použitého modelu. Modely zahrnující fixní efekty (*fixed effects*) mívají menší stabilitu než modely, které je nezahrnují. Autoři proto doporučují zvětšit zkoumané vzorky a omezit zahrnování fixních efektů do modelů.

Výsledky výzkumů snažících se odpovědět na obecnější výzkumné otázky typu: *Jaký je rozdíl mezi přidanou hodnotou zkušených a méně zkušených učitelů, případně: Jaké obecné změny ve výsledcích žáků lze pozorovat poté, co jim byl přiřazen učitel s naměřenou vyšší přidanou hodnotou*, nejsou příliš závislé na tom, jaká byla naměřená přidaná hodnota konkrétního učitele, a jak se měnila z roku na rok. Možná nestabilita přidané hodnoty jednotlivých učitelů totiž v takových případech neovlivní celkové výsledky studie. Zásadně odlišná je ovšem situace, pokud mají být výsledky měření přidané hodnoty učitelů použity k nějakým rozhodnutím, které se mají týkat jednotlivých učitelů. V tomto případě je potenciální nestabilita měření přidané hodnoty učitele podstatně větším problémem, protože by chyby měření mohly mít neoprávněně velký vliv na výši jeho finančních odměn nebo dokonce na jeho profesní kariéru. Provedené empirické studie nicméně potvrzují, že míra nestability u měření

přidané hodnoty učitele nebývá zpravidla tak vysoká, aby zabránila smysluplnému použití modelových odhadů přidané hodnoty učitele v praxi.

1.5.7 Velikost účinku přidané hodnoty učitele a délka jeho působení na žáky

Bacher-Hicks a Koedel (2023) se ve svých výzkumech soustředili i na problematiku velikosti účinku (effect size), což je hodnota měřící sílu vztahu mezi dvěma proměnnými v populaci, a také na odhadování této veličiny na základě výběrového vzorku. Velikost účinku se v matematické statistice definuje jako míra síly jevu, která je nezávislá na jednotkách měření. K běžně používaným statistikám tohoto typu patří míry statistické závislosti

Tabulka 1: Odhadovaná směrodatná odchylka efektivity učitele měřená ve směrodatných odchylkách výsledků testů studijních výsledků žáků

Study	Location	Teacher Effectiveness	
		Reading	Math
Rockoff (2004)	New Jersey	0.10	0.11
Nye, Konstantopoulos, and Hedges (2004)	Tennessee	0.26	0.36
Rivkin, Hanushek, and Kain (2005)	Texas	0.10	0.11
Aaronson, Barrow, and Sander (2007)	Chicago	—	0.13
Kane, Rockoff, and Staiger (2008)	New York City	0.08	0.11
Jacob and Lefgren (2008)	Undisclosed city	0.12	0.26
Kane and Staiger (2008)	Los Angeles	0.18	0.22
Koedel and Betts (2009)	San Diego	—	0.23
Jesse Rothstein (2010)	North Carolina	0.11	0.15
Hanushek and Rivkin (2010)	Undisclosed city	—	0.11

Poznámky: Všechny odhady udávají směrodatnou odchylku efektivity učitele měřenou prostřednictvím směrodatné odchylky testových výsledků žáků standardizovaných na střední hodnotu 0 a rozptyl 1. Všechny odchylky jsou opraveny o chybu měření v testu a s výjimkou Kanea a Staigera (2008) jsou odhadovány v rámci jednotlivé školy po ročních (within school-by-year) nebo v rámci školy po třídách a po ročních (within school-by-grade-by-year). Tato tabulka je převzata z článku Hanusheka a Rivkina (2010). Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. American economic review, 100(2), 267–271.

spojitých proměnných (např. korelační koeficient či koeficient determinace), míry závislosti nominálních znaků (např. Pearsonův koeficient ϕ nebo Cramerovo V) či míry rozdílu průměrů (např. Cohenovo d či Hedgesovo g). Velikost účinku má tu výhodu, že umožňuje snadno posoudit praktickou významnost statistické závislosti. V praxi modelování přidané hodnoty se zpravidla počítá jak velký efekt na změny ve výsledcích testu žáka má změna přidané hodnoty učitele o jednu směrodatnou odchylku jejich rozložení. Hanushek a Rivkin (2010) ve své přehledové studii uvádějí, že posun o jednu směrodatnou odchylku v přidané hodnotě učitele změní výsledky žáků ve standardizovaných srovnávacích testech matematiky o 0,11–0,36 směrodatné odchylky a o 0,08–0,26 směrodatné odchylky v testech čtení. Další empirické studie uvádějí užší rozmezí velikost účinku: 0,10–0,15 směrodatné odchylky u testů z matematiky a 0,05–0,15 směrodatné odchylky u testů ze čtení (viz tabulka 1).

Řada vědců se následně zaměřila na snahu vysvětlit, proč je efekt přidané hodnoty učitele větší u testů z matematiky než u testů ze čtení. Jedním z vysvětlení je působení efektů mimo školu, tedy především vliv rodičů, kteří jsou schopni u svých dětí více ovlivnit výkony ve čtení než výkony v matematice. Pokud jsou porovnávány velikosti účinku u přidané hodnoty učitelů 1. stupně základních škol a 2. stupně základních škol, opakovaně se ukazuje, že u učitelů 1. stupně základních škol lze zaznamenat větší variabilitu v působení přidané hodnoty učitelů na výsledky žáků, než je tomu v případě učitelů 2. stupně základních škol. To se vysvětluje jednak tím, že žáci na 1. stupni základních škol jsou tvárnější, jednak tím, že učitelé na 1. stupni základních škol se skupinou jim svěřených žáků obecně tráví více času.

Několik studií bylo věnováno zkoumání toho, jak dlouhodobý je vliv přidané hodnoty učitele. V praxi to zpravidla znamená, že výzkumníci několik let po sobě měří výsledky žáků poté, co začali být vystaveni působení učitele s velmi vysokou (nebo velmi nízkou) přidanou hodnotou. Studie ukazují, že vliv učitele se časem zmenšuje. Jednou naměřený efekt působení přidané hodnoty učitele se po dalším roce jeho působení zmenší o zhruba 50 procent. Panelová studie Chetty a kol. (2014a), která studovala vliv učitele v dlouhodobějším pohledu, ukázala, že vliv učitele se časem zmenšuje, až po cca 3–4 letech dosáhne 20–25 procent původních hodnot a poté už dále neklesá.

1.6 Jak souvisí přidaná hodnota učitele s jeho dalšími charakteristikami

Modely, pomocí nichž se může měřit přidaná hodnota učitele, ukazují, že učitelé ve velikosti jejich přidané hodnoty podstatně liší. Nabízí se tedy otázka, proč tomu tak je. S cílem odpovědět na tuto výzkumnou otázku se rozvinul výzkumný směr, který se snaží hledat souvislosti mezi naměřenou přidanou hodnotou učitele a jeho dalšími měřitelnými charakteristikami. Kromě teoretického významu mají tato zkoumání i význam praktický, protože umožňuje identifikovat charakteristiky učitelů, které mohou mít souvislost s přidanou hodnotou učitelů a na základě znalosti těchto charakteristik učitelů jejich pedagogickou efektivitu predikovat.

1.6.1 Souvislost mezi přidanou hodnotu učitele a jeho osobními charakteristikami

Bacher-Hicks a Koedel (2023) se ve své studii věnují problematice vztahu mezi přidanou hodnotou učitele a jeho dalšími osobními charakteristikami. Uvádají, že se v odborné literatuře při zkoumání přidané hodnoty

učitele velmi často pracuje s jeho kvalifikací a dalšími charakteristikami, o kterých se předpokládá, že mají přímou souvislost s velikostí jeho přidané hodnoty. Za nejčastěji uvažované charakteristiky autoři studie považují demografické údaje učitele (věk, pohlaví), délku jeho praxe, dosažené vzdělání, specializaci/obor studia, dosažené skóre v certifikačních testech, prestiž vystudované školy atd. Na základě rešerše literatury ovšem docházejí k závěru, že sledované charakteristiky učitelů zpravidla vysvětlují jen malou část variance v přidané hodnotě učitelů (Aaronson a kol., 2007).

Výzkumy naproti tomu ukazují, že ze sledovaných charakteristik mají na efektivitu učitelů konsistentně největší vliv jejich zkušenosti. Bylo zjištěno, že zkušenější učitelé mívají v průměru vyšší přidanou hodnotu. To je vysvětlováno tím, že v průběhu praxe získávají důležité zkušenosti, ale zlepšují si i svoje pedagogické dovednosti. To platí zejména v raných fázích jejich kariéry. Studie, které se zaměřují na měření vývoje přidané hodnoty učitele, se shodují v tom, že na počátku kariéry se pedagogické schopnosti učitele (a tedy i jejich přidaná hodnota) zvyšují rychle, ovšem po nějaké době se zvyšovat přestanou. Neexistuje ovšem úplná shoda na tom, kdy se přestane přidaná hodnota učitelů zvyšovat, většina studií tvrdí, že již po několika letech po vstupu do praxe. Kromě zlepšování pedagogických schopností učitelů v průběhu jejich kariéry existuje ještě alternativní vysvětlení, proč je naměřená přidaná hodnota učitele systematicky vyšší u učitelů s delší praxí. Vysvětlení je založeno na myšlence, že učitelé s nižší přidanou hodnotou nejsou ve své profesi příliš úspěšní, a proto ze škol odcházejí s větší pravděpodobností než úspěšní učitelé s vyšší přidanou hodnotou. Výsledkem tohoto procesu může být pozorovaná skutečnost, že mezi zkušenými učiteli přirozeně stoupá průměrná přidaná hodnota učitele. Obě vysvětlení se se vzájemně nevylučují, ale ve skutečnosti se mohou doplňovat, protože oba výše popsane jevy lze v praxi pozorovat.

1.6.2 Souvislost mezi přidanou hodnotu učitele a jinými indikátory jeho výkonu

Měření přidané hodnoty učitele pomocí modelů postavených na analýze výsledků jejich žáků ve srovnávacích testech není jediným možným způsobem, jak lze výkon učitele měřit. Bacher-Hicks a Koedel (2023) proto věnovali pozornost také studiím, které zkoumaly vztah mezi odhadovanou přidanou hodnotou učitele a dalšími měřitelnými charakteristikami jeho výkonu. Upozorňují, že tyto další indikátory výkonu učitele mohou zahrnovat např. hodnocení výkonu učitele ředitelem, výsledky z hospitací (expertů) nebo informace o kvalitě výuky získané z dotazníkového šetření realizovaného mezi žáky školy. Harris a Sass (2014) a Jacob a Lefgren (2008) realizovali korelační studie, ve kterých se snažili identifikovat souvislost mezi modelovými odhady přidané hodnoty učitelů a jejich hodnocením řediteli příslušných škol. Studie obou dvojic autorů ukázaly, že ředitelé jsou schopni s velkou mírou přesnosti identifikovat učitele nacházející se na obou koncích spektra přidané hodnoty (tedy nejlepší a nejhorší učitele), ale s mnohem menší přesností dokážou odhadnout rozdíly mezi učiteli, kteří se nacházejí ve střední části spektra. Korelační koeficienty mezi modelově odhadovanými přidanými hodnotami učitele a hodnocením, které jim přidělili jejich ředitelé, se pohybovaly v rozmezí 0,29–0,44.

Nejrozsáhlejší výzkum v této oblasti byl proveden v rámci studie *Measures of Effective Teaching* (MET)¹², která shromáždila videa ukazující didaktickou práci více než 1 000 učitelů v jejich třídách v šesti různých školských distriktech. Následně vyškolení externí hodnotitelé použili pět různých instrumentů, aby ohodnotili výkon učitelů v pořízených nahrávkách. Kromě toho byli požádáni žáci hodnocených učitelů, aby se účastnili dotazníkového šetření a ohodnotili klima ve třídě a kvalitu výuky poskytnuté učitelem. Kane a Steiger (2012) následně využili data o výsledcích žáků ve standardizovaných testech k modelovému odhadu přidané hodnoty učitelů. Nakonec porovnali výsledky hodnocení práce učitelů prostřednictvím evaluace videí a modelů přidané hodnoty a došli k závěru, že výsledky obou měření pozitivně korelují, u matematiky mezi 0,16–0,26. Korelace mezi modelovými odhady přidané hodnoty a evaluací práce učitelů jejich žáky byla vyšší, hodnota korelačního koeficientu dosáhla 0,37. Pozitivní korelace mezi modelovými odhady přidané hodnoty a jinými mírami efektivnosti práce učitele dávají důvěryhodnost modelovým odhadům. Hodnoty korelačních koeficientů jsou však relativně nízké, protože:

- (a) modelové odhady mají určitou míru nepřesnosti,
- (b) alternativní míry hodnocení výkonu učitele mohou být zatíženy neznámou systematickou chybou,
- (c) alternativní míry hodnocení výkonu učitele mohou zachycovat některé dimenze práce učitele, které nezachycují modelové odhady přidané hodnoty učitele.

Obecně přitom platí, že alternativní míry hodnocení výkonu učitele jsou schopny predikovat výsledky žáků, ale ne tak dobře, jak jsou toho schopné modelové odhady přidané hodnoty.

1.7 Rozšíření klasických modelů

V současné době lze v odborné literatuře nalézt pokusy o rozšiřování používání klasických modelů přidané hodnoty (Bacher-Hicks, Koedel, 2023). Bacher-Hicks a Koedel (2023) udávají, že se modely měřící přidanou hodnotu v současné době používají nejen na základních školách, ale také na středních školách, zároveň vznikají pokusy o odvozování přidané hodnoty učitelů nejen od kognitivních výsledků jejich žáků, ale také od nekognitivních výsledků. Modely přidané hodnoty se také začínají používat i v jiných zemích než pouze ve Spojených státech a expandují i do neučitelských profesí, například se používají k hodnocení efektivity ředitelů škol, ale také školských politik apod.

¹² Projekt MET je výzkumným partnerstvím akademiků, učitelů a vzdělávacích organizací, které se zavázaly hledat lepší způsoby, jak identifikovat a rozvíjet efektivní výuku. Financování tohoto projektu zajišťovala *Nadace Billa a Melindy Gatesových*. Na projektu spolupracovalo přibližně 3 000 dobrovolníků z řad učitelů ze šesti distriktů. Mezi partnery projektu byly desítky odborníků na vzdělávání a výzkumných pracovníků z univerzit a další institucí a organizací. Projekt byl navržen s cílem zjistit, jak nejlépe využít metody hodnocení k tomu, aby učitelům sdělily více informací o dovednostech, které je činí nejefektivnějšími, a jak pomoci distriktům identifikovat a rozvíjet skvělou výuku. Informace, které byly shromážděny mezi roky 2010–13 umožnily mnohem lepší pochopení toho, jak vypadá skvělá výuka. Zároveň tyto poznatky pomohly vytvořit nástroje a systémy, které pomáhají učitelům pomoci žákům uspět. Na základě zkušeností získaných v rámci projektu MET bylo určeno devět hlavních principů, které budou sloužit jako podklad pro návrh a implementaci vysoce kvalitních systémů podpory a hodnocení učitelů, což v konečném důsledku povede k lepšímu výkonu žáků.

1.7.1 Modely měřící přidanou hodnotu učitelů na středních školách

První rozšíření představují modely, které se zaměřují na vliv přidané hodnoty učitele na výsledky žáků středních škol (v USA *high schools*, ročníky 9–12), viz např. studie Aaronsona a kol. (2007). Původně se totiž většina modelů přidané hodnoty totiž zabývala vlivem přidané hodnoty učitele na žáky nižších ročníků škol, typicky na žáky základních škol. Americké střední školy se ovšem v mnohém odlišují od základních škol, zpravidla jsou větší a žáci v nich netvoří stálé třídy, ale jsou v každém předmětu rozdělováni do různých tříd, protože se předměty i rozvrh hodin přizpůsobuje úrovni jejich znalostí a jejich individuálním zájmům. Zatímco v nižších stupních škol jsou všechny děti ve třídě spolu a na jednotlivé předměty mají stejné učitele, výuka na středních školách je velmi personalizovaná, takže žáci navštěvující stejný ročník mohou být např. v matematice umístěni do odlišných tříd a mohou matematiku studovat na různé úrovni. Všechny čtyři výše citované studie zkoumající vliv přidané hodnoty učitelů na výsledky testů žáků, pracovaly v modelech s fixními efekty na úrovni školy a to buď standardními fixními efekty definovanými na úrovni celé školy, nebo fixními efekty škola-kurz, či škola-specifický set kurzů. Všechny modely se musely potýkat nejen s problémem nenáhodného přiřazování žáků školám, ale i s problémem nenáhodného přiřazování učitelů kurzům v rámci jednotlivých škol. Je obecně mnohem těžší zachytit vliv přidané hodnoty učitele na výsledky žáků v prostředí, kde mají žáci de facto individualizovaný studijní plán, a proto i individualizovaný soubor vyučujících. Jednotliví učitelé navíc mohou učit ve stejném předmětu různé žáky na různých úrovních. Zatímco jeden ze žáků je vyučován základům, jiný žák téhož ročníku je vyučován stejným učitelem vyšší úroveň matematiky. Jeden a tentýž učitel proto také nemusí mít stejnou přidanou hodnotu, dokonce ani ne pro žáky stejného předmětu ve stejném ročníku. Výsledky studií nicméně obecně potvrzují, že existuje vliv přidané hodnoty učitele na výsledky žáků. Změna přidané hodnoty učitele o jednu směrodatnou odchylku vedla ke změně výsledků testů o 0,08–0,21 v matematice a o 0,03–0,10 v předmětu anglický jazyk.

Vzhledem k malému počtu studií provedených na středních školách ovšem nelze jednoduše porovnat tyto výsledky s výsledky zjištěnými na základních školách. Zdá se ovšem, že na středních školách je efekt přidané hodnoty nižší. Autoři to vysvětlují třemi možnými způsoby: (a) přidaná hodnota učitele na středních školách je substantivně odlišná od přidané hodnoty učitele v základním vzdělávání, (b) učitelé na středních školách mají obecně menší vliv na žáky a (c) na středních školách má obecně menší význam používat výsledky žáků ve srovnávacích testech jako měřítko hodnocení práce jejich učitelů.

1.7.2 Modely

Druhou oblastí rozšíření klasických modelů přidané hodnoty je odvozovat přidanou hodnotu učitelů od jiných výsledků než výsledků jejich žáků v testech (Bacher-Hicks, Koedel, 2023). V odborné literatuře se někdy hovoří o nekognitivních výsledcích, ačkoliv nepanuje plná shoda na tom, jaký výsledek lze považovat za kognitivní a jaký nikoliv. Obecně lze za jiné výsledky než testové označit například: školní známky, docházka, podmíněčné vyloučení, opakování ročníku, motivace k vlastnímu rozvoji, vytrvalost, úsilí, pocity štěstí a chování ve třídě. Metodologicky jsou takto zaměřené studie podobné klasickým, jen místo výsledků standardizovaných testů se jako analogická závisle proměnná použije nějaká agregátní proměnná vytvořená jako index z některých výše uvedených ukazatelů.

Provedené studie (např. Liu, Loeb 2021) ukazují, že učitelé mají vliv na netestové charakteristiky typu známky, opakování ročníku či docházka, a velikost tohoto vlivu odpovídá vlivu na výsledky v testech. Kraft (2019) prokázal vliv přidané hodnoty učitelů na motivace k vlastnímu rozvoji, na vytrvalost a úsilí, které jsou v literatuře také označovány jako socio-emocionální kompetence. Slabým místem jeho výzkumu byla skutečnost, že model vytvářel na datech od pouhých 236 učitelů. Při modelování byl kontrolován vliv řady dalších proměnných, jako byly skóre v testech, informace o žácích získané dotazníkovým šetřením rok před provedením vlastní studie, standardní demografické a sociálně-ekonomické charakteristiky žáků. Kraft prokázal vliv přidané hodnoty učitelů na sledovaný soubor závisle proměnných. Velikost tohoto vlivu byla opět podobná vlivu na výsledky žáků ve standardizovaných testech z matematiky a čtení. Nicméně byla nalezena jen mírná korelace mezi vlivem učitelů na výsledky v testech a netestové charakteristiky, takže jedním ze závěrů studie bylo, že výsledky žáků ve standardizovaných testech nelze jednoduše nahradit využitím netestových ukazatelů. Blazar (2018) a Blazar a Kraft (2017) analyzovali vliv přidané hodnoty učitelů na chování žáků ve třídě, víru ve vlastní schopnosti (*self-efficacy*) v matematice a subjektivně pociťované štěstí (všechny uvedené údaje byly získané na základě self-reportingu žáků). Výsledky studií prokázaly, že přidaná hodnota učitele má vliv na socio-emocionální charakteristiky žáků, vliv je zhruba stejně silný jako vliv na výsledky v testech, ale socio-emocionální charakteristiky a výsledků testů nejsou vzájemně příliš těsně korelovány.

1.7.3 Modely měřící přidanou hodnotu učitelů v jiných zemích

Další oblastí, ve které byly klasické modely odhadu přidané hodnoty učitelů rozšířeny a následně upravovány, byly analýzy provedené na datech z jiných zemí (Bacher-Hicks, Koedel, 2023). Autoři tuto skutečnost považují za významnou, a to především proto, že tím byly znovu testovány některé z předpokladů, které se používaly v klasických modelech. Zatímco modely ve Spojených státech ukázaly, že zkreslení odhadu není příliš významně ovlivněna nenáhodným přidělováním žáků učitelům, výsledky analýz provedených v jiných zemích byly v tomto ohledu méně jasné. Zejména analýzy provedené v rozvojových zemích ukázaly, že vliv nenáhodného přidělování učitelů žákům je minimálně stejný, ale zpravidla větší, než byl pozorován v USA. Studie z pakistánského Pandžábu (Bau, Das 2020) ukázala, že zvýšení přidané hodnoty učitele o jednu směrodatnou odchylku zvýší výkon žáků měřeného standardizovaným testem v průměru o 0,15 směrodatné odchylky. Demografické a socio-ekonomické charakteristiky učitelů, stejně jako v USA, nebyly příliš dobrými prediktory výsledků žáků v testech. Na rozdíl od USA byla ovšem významným prediktorem výsledků žáků v testech úroveň znalostí obsahu vyučovaného předmětu jejich učitelem. Vysvětlením byla obecně větší variabilita v úrovni znalostí obsahu mezi pakistánskými učiteli ve srovnání s jejich americkými kolegy.

Studie, kterou realizovali Araujo a kol. (2016) v Ekvádoru, je v mnohém metodologicky originální. Autoři studie totiž dostali možnost přiřadit v mateřských školách 24 000 dětí do jednotlivých tříd kvazi-náhodným způsobem, podle abecedního pořadí jejich příjmení. Proto byla studie vzácnou kombinací náhodného přiřazování žáků učitelům a využití velkého vzorku testovaných dětí. Výsledky studie ukázaly, že zvýšení přidané hodnoty učitele o jednu směrodatnou odchylku zvýšilo výsledky testu dětí v matematice a jazycích o 0,09 standardní odchylky.

V severní Ugandě provedli obdobnou studii Buhl-Wiggers a kol. (2022), kteří dostali možnost přiřadit náhodným způsobem žáky a učitele ve vybraných základních školách. Zvýšení přidané hodnoty učitele o jednu směrodatnou odchylku zvýšilo výsledky testu ve čtení o 0,18 standardní odchylky. Azam a Kingdom (2015) ve studii provedené na deseti středních školách v Indii identifikovali silnější efekt přidané hodnoty učitele – její zvýšení o směrodatnou odchylku zvýšilo po dvou letech studia výsledky v testech o 0,37 směrodatné odchylky. Crawford a Rolleston (2020) odhadli analogická čísla pro Vietnam na 0,28 a pro Etiopii na 0,30 směrodatné odchylky. Autoři kapitoly ve svém přehledu zmiňují jenom dvě studie provedené ve vyspělých zemích mimo USA. Slater a kol. (2012) odhadli vliv přidané hodnoty učitelů ve Velké Británii na 0,23 směrodatné odchylky ve výsledcích testů. Tanaka a kol. (2020) po analýze v základních a středních školách v jednom z velkých japonských měst odhadli vliv přidané hodnoty učitelů na 0,15 směrodatné odchylky u testů z jazyků a 0,18 směrodatné odchylky u testů z matematiky.

1.7.4 Modely měřící přidanou hodnotu v neučitelských profesích

Odborná literatura se rovněž zabývá měřením možného vlivu přidané hodnoty v neučitelských profesích (Bacher-Hicks, Koedel, 2023). Nejčastěji zkoumanou neučitelskou profesí jsou ředitelé škol. Odhadovat vliv přidané hodnoty ředitele je obecně podstatně složitější, než odhadoval vliv přidané hodnoty učitele. Metodologicky komplikované je především to, že:

(a) všichni žáci jedné školy mají jediného ředitele,

(b) je obecně mnohem menší počet ředitelů než učitelů a ředitelé mají v průměru kratší dobu kariéry na svém postu, než je průměrná doba kariéry učitelů na jejich pozicích.

Vzhledem k tomu, že všichni žáci jedné školy mají jediného ředitele, je obtížné odlišit vliv ředitele jako konkrétní osoby od obecného efektu školy. Ředitelé ovšem mohou do určité míry ovlivnit průměrnou kvalitu učitelů, a tím mít vliv na výsledky žáků. Modely odhadující vliv ředitele na výsledky žáků jsou zpravidla postaveny na myšlence porovnávat výsledky žáků ve stejné škole ve dvou různých obdobích, před a po změně ředitele. Určitý problém je ovšem v tom, že vliv ředitele se nemusí projevit dostatečně rychle na to, aby byl zaznamenán během doby studia jedné kohorty žáků. V prvním roce po svém nástupu zpravidla nový ředitel „zdědí“ většinu učitelského sboru, který ve škole vyučoval za předchozího ředitele, a také klíčové školní politiky, kterými se řídí chod školy. I pokud nový ředitel věci ve škole cíleně mění, vliv nového ředitele na učitelský sbor i chod školy se začne projevovat až po nějaké době. Druhý problém spočívá v tom, že ředitelé obecně nemění svoje působíště příliš často a zejména nepřecházejí ve velkém počtu z jedné školy do druhé. Výzkumníci zajímající se o měření přidané hodnoty ředitele výše popsaným způsobem tedy mívají potíže s velikostí vzorku použitelných případů.

I.8 Využití přidané hodnoty učitelů v politikách¹³

Výzkumům přidané hodnoty učitelů je věnována významná pozornost zejména proto, že se jedná o potenciálně aplikovatelnou výzkumnou problematiku. Z výzkumů obecně vyplývá, že není pochyb o tom, že existují významné rozdíly v efektivitě učitelů, a že přístupy vedoucí ke zlepšení kvality učitelů by mohly dramaticky ovlivnit výsledky žáků. Hanushek (2009) ukazuje, že odstranění 6–10 procent nejhorších učitelů, by mohlo mít silný pozitivní dopad na výsledky žáků, a to i v případě, že by tito učitelé byli trvale nahrazeni „pouze“ průměrnými kolegy.

Větší problémy ovšem nastávají, pokud by odhady efektivity učitelů, založené na jejich přidané hodnotě měly sloužit jako podklad k rozhodování o výši finančního ohodnocení učitelů, o jejich zaměstnání, povýšení nebo rozřazení do pozic ve škole. Možnost zavedení odměňování učitelů za výkon, a to na základě jejich odhadované přidané hodnoty, je sice motivátorem řady výzkumů v dané oblasti, ale vývoj školských politik tímto směrem je zatím velmi omezen (Podgursky, Springer 2007), protože panují obavy z možných negativních dopadů na školy, učitele i žáky (Hanushek, Rivkin, 2010). Využívání odhadů přidané hodnoty učitelů v praxi má totiž řadu úskalí, které byly diskutovány výše, ale i rizik, které by mohly vyplynout později, například zavedení měření přidané hodnoty učitelů by mohlo vést k podvádění, zaměřování obsahu výuky výlučně na testovanou problematiku a ignoraci obsahů a předmětů, které testování nepodléhají apod.

Ačkoli výzkumníci mohou zmírnit dopady výběrové chyby na odhady přidané hodnoty učitelů, chyba by nevyhnutelně mohla vést k tomu, že by někteří úspěšní učitelé dostali nízké hodnocení, a naopak někteří neúspěšní učitelé vysoké hodnocení. Problémy s chybovostí v měření by do značné míry zmizely, pokud by učitelé byli pozorováni dlouhodobě v rozsahu několika let a při velkém počtu jimi vyučovaných dětí (McCaffrey a kol. 2009). Spoléhání se na longitudinální data ovšem vylučuje ze systému nové učitele a tlumí sílu pobídek, protože pracovní výkon v aktuálním roce by jen částečně určoval míru efektivity.

Je potřeba zvážit i problematiku spravedlnosti, jakékoli nezohlednění vlivu přidělování žáků do tříd podle nepozorovaných (a neměřených) charakteristik by mohlo penalizovat učitele s obtížněji vzdělávanými školními třídami, a naopak odměňovat učitele s méně obtížně vzdělávanými školními třídami, aniž by použité metody byly schopny odhalit vliv rozdílů mezi vyučovanými třídami. To by mohlo odrazovat od rozhodnutí, která mohou být obecně pro kvalitu poskytování vzdělávání prospěšná, včetně například zařazování hůře vzdělávaných nebo rušivých žáků ke kvalitnějším učitelům. Tato potenciální nevýhoda však může být zmírněna kombinací subjektivního hodnocení supervizora nebo kolegů s objektivními odhady přidané hodnoty, protože ředitelé by mohli odhady zasadit do kontextu a zdá se, že jsou schopni posoudit rozdíly v účinnosti přinejmenším na obou koncích statistického rozdělení přidané hodnoty učitelů (Jacob, Lefgren, 2008).

Konečně, soustředit se na rozdíly v rámci školy nemusí být pro politiku vhodné. Zaměření na rozdíly existující v rámci školy, aplikované kvůli obtížím při zohlednění rozdílů mezi školami, vyvolává obavy ohledně hodnocení výkonu; některé školy mohou mít v průměru mnohem lepší učitele než jiné a bylo by důležité takové rozdíly rozpoznat a vzít v potaz.

¹³ Tato část textu je podrobným výtahem z článku Hanusheka a Rivkina (2010). Hanushek, E. A., Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American economic review*, 100(2), 267–271. Z tohoto textu je přebírán i odkazový aparát.

Potenciální problémy naznačují, že statistické odhady efektivity učitele založené na výsledcích jeho žáků v testech čtení a matematiky by neměly tvořit jedinou složku jakéhokoli systému hodnocení. I tak je klíčovou politickou otázkou, zda opatření s přidanou hodnotou, navzdory svým nedostatkům, mohou poskytnout cenné informace pro zlepšení personálních rozhodnutí, která v současnosti vycházejí z omezených informací o efektivitě učitelů. Důvody pro objektivní opatření jsou pravděpodobně nejsilnější v těch městských nebo venkovských oblastech, kde existuje omezenější konkurence mezi veřejnými a soukromými školami. V takových místech může být velkým příslibem hybridní přístup k hodnocení, v němž opatření s přidanou hodnotou tvoří jednu z několika jeho složek.

I.9 Literatura

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95–135.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, 131(3), 1415–1453.

ASA (2014). *ASA Statement on Using Value-Added Models for Educational Assessment*. American Statistical Association. <https://web.archive.org/web/20170925023105/http://www.amstat.org/asa/files/pdfs/POL-ASA-VAM-Statement.pdf>.

Azam, M., & Kingdon, G. G. (2015). Assessing teacher quality in India. *Journal of Development Economics*, 117, 74–83.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73, 101919.

Bacher-Hicks, A., Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. Pp. 93–134 in Hanushek, Eric A., Machin, S., Woessmann, L. (eds.) *Handbook of the Economics of Education*. Volume 6. San Diego: Elsevier Science & Technology.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... & Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278. *Economic Policy Institute*. <https://eric.ed.gov/?id=ED516803>

Bau, N., & Das, J. (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, 12(1), 62–96.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, 13(3), 281–309.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational evaluation and policy analysis*, 39(1), 146–170.

Bracey, G. (2013). *Value Subtracted: A "Debate" with William Sanders*. Huffpost. https://www.huffpost.com/entry/value-subtracted-a-debate_b_47404.

Braun, H. I. (2005). Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models. Policy Information Perspective. *Educational Testing Service*. <https://www.ets.org/Media/Research/pdf/PICVAM.pdf>

Buhl-Wiggers, J., Kerwin, J. T., Smith, J., & Thornton, R. (2022). Learning More about Teachers: Estimating Teacher Value-Added and Treatment Effects on Teacher Value-Added in Northern Uganda. https://jasonkerwin.com/Papers/TVA/Buhl-Wiggers_Kerwin_Smith_and_Thornton_TVA_Latest.pdf

Crawford, L., & Rolleston, C. (2020). Long-run effects of teachers in developing countries. *Review of Development Economics*, 24(4), 1279–1299.

Dillon, S. (2010). *Method to grade teachers provokes battles*. New York Times, 31. <https://ed.stanford.edu/in-the-media/method-grade-teachers-provokes-battles>

Ewing, J. (2011). *Mathematical Intimidation: Driven by the Data*. NOTICES OF THE AMS <https://web.archive.org/web/20170805021551/http://www.ams.org/journals/notices/201105/rtx110500667p.pdf>.

Fernando, J. (2023). *What you need to know about the two basic types—traditional and Roth*. Investopedia. <https://www.investopedia.com/terms/I/401kplan.asp>.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, 80(319), 589–612.

Green, E. (2014). Building a better teacher: How teaching works (and how to teach it to everyone). W. W. Norton & Company; Illustrated edition (August 4, 2014). <https://www.amazon.com/Building-Better-Teacher-Teaching-Everyone-ebook/dp/B00FPT5MSQ>

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 61(2), 280–288.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of political economy*, 100(1), 84–117.

Hanushek, E. A. (2009). Teacher Deselection. In *Creating a New Teaching Profession*, edited by Dan Goldhaber

and Jane Hannaway (165–80). Washington, DC: Urban Institute Press.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American economic review*, 100(2), 267–271.

Hanushek, E. A., and S. G. Rivkin. (2006). Teacher Quality. In *Handbook of the Economics of Education*, edited by Eric A. Hanushek and Finis Welch (1051–78). Amsterdam: North Holland.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Disruption versus Tiebout improvement: The costs and benefits of switching schools. *Journal of public Economics*, 88(9–10), 1721–1746.

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, 40, 183–204.

Hofer, S. I., Holzberger, D., & Reiss, K. (2020). Evaluating school inspection effectiveness: A systematic research synthesis on 30 years of international research. *Studies in Educational Evaluation*, 65, 100864. <https://www.sciencedirect.com/science/article/abs/pii/S0191491X20301127>

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9), 2593–2632.

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Discussion of the American Statistical Association's Statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy*, 1(1), 111–113.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education finance and Policy*, 4(4), 520–536.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 101–136.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research. <https://www.nber.org/papers/w14607>

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. *Policy and practice brief prepared for the Bill and Melinda Gates Foundation*. <https://eric.ed.gov/?id=ED540960>

Koedel, C., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47, 180–195.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, 54(1), 1–36.

Liu, J., & Loeb, S. (2021). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, 56(2), 343–379.

McCaffrey, D. F., & Lockwood, J. R. (2008, November). Value-added models: Analytic issues. In *A workshop held by the Committee on Value-added Methodology for Institutional Improvement, Program Evaluation and Educational Accountability sponsored by the National Research Council and the National Academy of Education, Washington, DC*.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education finance and Policy*, 4(4), 572–606.

Murnane, R. J. (1974). *The impact of school resources on the learning of inner-city children*. Yale University. <https://eric.ed.gov/?id=ED121905>

National Research Council. (2010). *Getting Value Out of Value-Added: Report of a Workshop*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/12820>.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education policy analysis archives*, 18, 23–23.

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, 51(2), 328–362.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of policy analysis and management*, 26(4), 909–949.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 89(2), 146–155.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education finance and policy*, 4(4), 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1), 175–214.

Ryška, R. (2008). *Modelování přidané hodnoty ve vzdělávání: Shrnutí*. Výstup projektu LS0602: Zjišťování přidané hodnoty. Univerzita Karlova v Praze, Pedagogická fakulta. <http://www.strediskovzdelavacipolitiky.info/download/VAM%20shrnuti%202008.pdf>

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Research Progress Report. University of Tennessee Value-Added Research and Assessment Center. <https://www.beteronderwijsnederland.nl/files/cumulative%20and%20residual%20effects%20of%20teachers.pdf>

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of personnel evaluation in education*, 11, 57–67.

Slater, H., Davies, N. M., & Burgess, S. (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics*, 74(5), 629–645.

Tanaka, R., Bessho, S. I., Kawamura, A., Noguchi, H., & Ushijima, K. (2020). Determinants of teacher value-added in public primary schools: Evidence from administrative panel data. *IZA Discussion Paper No. 13146* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3579238

Tierney, R. J., Rizvi, F., & Ercikan, K. (2022). *International Encyclopedia of Education, Fourth Edition*. Elsevier Science. ISBN: 9780128186305

2. Modelování přidané hodnoty – CFR-I a CFR-II

Chetty, Friedman a Rockoff (2014a, 2014b) si položili otázku jak měřit a zlepšovat kvalitu výuky na základních školách. Konkrétně je zajímalo, zda je vliv učitelů na výsledky testů jejich žáků (tzv. „přidaná hodnota učitele“) dobrým měřítkem kvality učitelů. Jedním z důvodů, proč tato otázka vyvolala diskusi, je neshoda ohledně toho, zda zjišťování a měření přidané hodnoty učitelů poskytují nezkrácené odhady kauzálních efektů práce učitelů na výsledky žáků. Odpovědi na tuto otázku hledali Chettyho a kol. (2014a) ve své první studii. Chettyho a kol. (2014a) v této studii testovali, zda nedochází ke zkreslení měření přidané hodnoty učitele. K tomu účelu zahrnuli do modelů přidané hodnoty i dříve nepozorované charakteristiky rodičů, ale také kvazi-experimentální design založený na využití přirozených obměn učitelského sboru. Na základě administrativních záznamů ze školních distriktů a záznamů o více než jednom milionu žáků zjistili, že jejich modely přidané hodnoty učitelů, které kontrolují výsledky předchozích testů žáků, vykazují pouze malé zkreslení při předpovídání dopadů práce učitelů na výsledky žáků v testech (v odborných textech bývají měření uváděná v první studii autorů označována CFR-I)¹⁴.

Na první studii navázala druhá studie Chettyho a kol. (2014b). Autoři se v ní pokoušeli zjistit, zda je vliv učitelů na žáky dlouhodobý, či pouze krátkodobý. Konkrétně je zajímalo, zda mají učitelé vliv i na životní a pracovní dráhy svých žáků v pozdějším věku. Kvůli nedostatku empirických důkazů se jednalo o doposud nevyřešenou otázku, která vyvolávala velké debaty. Předmětem debaty bylo zejména, zda učitelé s vysokou přidanou hodnotou zlepšují dlouhodobé výsledky svých žáků či pouze zlepšují výsledky žáků ve standardizovaných testech. K nalezení odpovědi autoři využili data ze školských distriktů¹⁵ a data z finančních úřadů, především údaje uvedené v daňových přiznáních pro federální daň z příjmů. Tato data se podařilo propojit pro více než jeden milion žáků z velkého městského školského distriktu na východním pobřeží USA. Analýza dat ukázala, že žáci, které učili učitelé s vysokou přidanou hodnotou, později s větší pravděpodobností navštěvovali kvalitnější vysoké školy, měli v mladém dospělém věku v průměru vyšší platy a bylo mezi nimi relativně méně dívek, které porodily děti už před dvacátým rokem věku. Analýza rovněž prokázala, že nahrazení učitele, jehož přidaná hodnota byla v dolních 5 procentech hodnoty ukazatele, učitelem s průměrnou přidanou hodnotou by nezanedbatelně zvýšilo výši budoucího celoživotního příjmu žáků. V odborných textech bývají měření uváděná ve druhé studii autorů označována CFR-II.

2.1 Míra zkreslení odhadů přidané hodnoty učitele¹⁶

Chetty a kol. (2014a) udávají, že jedním z významných, ale zároveň kontroverzních způsobů měření kvality učitelů je posuzování jejich efektivity na základě vlivu, který mají na výsledky žáků ve standardizovaných testech, což je forma hodnocení, která se běžně označuje jako „přidaná hodnota učitele“. Udávají, že v současné

¹⁴ Název je odvozen od jmen autorů.

¹⁵ Data obsahovala výsledky testů žáků a informace o jejich učitelích.

¹⁶ Tato kapitola je podrobným výtahem z textu: Chetty, R., Friedman, J. N., Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9), 2593–2632. Z tohoto textu je přebírán i odkazový aparát.

době je tento způsob hodnocení učitelů běžně používán v řadě školských distriktů po celých Spojených státech (od Washingtonu, D. C. po Los Angeles). Zároveň upozorňují, že tento přístup má jak své zastánce (např. Gordon, Kane a Staiger 2006, Hanushek 2009), kteří argumentují, že vybírání učitelů na základě jejich přidané hodnoty vede k podstatnému zlepšení výsledků žáků, ale i kritiky (např. Baker a kol. 2010, Corcoran 2010), kteří tvrdí, že míry přidané hodnoty učitelů jsou špatnými ukazateli jejich kvality. Přitom debaty o přínosech a limitech přidané hodnotě učitelů se točí zejména kolem dvou otázek. Zaprvé, zda rozdíly ve výsledcích testů žáků, které jsou základem pro měření přidané hodnoty učitele, skutečně zachycují kauzální efekty učitele, nebo zda se jedná o domnělé efekty zkreslené jinými faktory, například distribucí žáků do škol a tříd. Zadruhé, zda učitel, který zvyšuje výsledky žáků v testech, také zlepšuje jejich výsledky v pozdějších letech a dospělosti, nebo zda je prostě pouze lepší ve schopnostech připravit své žáky na úspěšné absolvování testů.

První studie – Chettyho a kol. (2014a) se zabývá první ze dvou výše uvedených otázek. Autoři studie upozorňují, že co se týče míry zkreslení přidané hodnoty učitele, předchozí výzkumy dospěly k rozporuplným závěrům (Kane a Staiger 2008, Rothstein 2010, Kane a kol. 2013). Přitom vyřešení tohoto odborného sporu považují za zásadní, a to především pro představitele decizní sféry, protože zkreslená měření přidané hodnoty by v praxi mohla vést k systematickému odměňování nebo penalizování učitelů nikoliv na základě jejich pedagogických schopností, ale na základě jiných charakteristik, např. v závislosti na složení žáků jimi vyučovaných tříd.

2.1.1 Data, metodologie, shrnutí výsledků

K tomuto účelu Chetty a kol. (2014a) vyvinuli nové metody pro výpočet míry zkreslení odhadů přidané hodnoty učitelů a testovali je na datech získaných ze dvou administrativních databází školských distriktů¹⁷. První databázi tvořil soubor dat, který obsahoval výsledky testů žáků třetích až osmých ročníků. Druhá databáze obsahovala informace o jejich učitelích.

Datový soubor, který obsahoval výsledky testů žáků, pocházel z velkého městského školského distriktu v USA. Data podávala informace o více než 2,5 milionech žáků a obsahovala výsledky z více než 18 milionů testů z matematiky a angličtiny zadávaných v letech 1989–2009. V průběhu dvacetiletého období, které autoři zkoumali, se však měnil režim testování. Na počátku sledovaného období byly všechny testy ve sledovaném školském distriktu specifické pouze pro tento distrikt. Od konce 90. let byly testy pro žáky čtvrtých a osmých ročníků základních škol (v souladu s federálním zákonem *No Child Left Behind*) zadávány v rámci celostátního systému testování a od roku 2006 byly tímto způsobem zadávány testy žákům již od třetích ročníků základních škol. Změnilo se i období, ve kterém probíhalo testování. V polovině 90. let se všechny testy zadávaly koncem dubna nebo v průběhu května. V následujících obdobích byly testy zadávány dříve, zpravidla již v průběhu února. Kvůli rozdílům v režimech testování postupovali autoři obdobně jako autoři některých předchozích studií a hodnoty skóre, které žáci v testech získali, standardizovali a přepočítali u každého testu v každém školním roce a v každém ročníku tak, aby byl průměr nula a směrodatná odchylka jedna. Kromě výsledků žáků v testech tento soubor obsahoval i informace o jejich etnické příslušnosti, pohlaví, věku, speciálních vzdělávacích potřebách

¹⁷ I když v samotné studii není uvedeno, ze kterých distriktů data pocházejí, v jiných textech (viz např. Rothstein, 2017a,b) je uvedeno, že se jedná o data ze škol v New Yorku.

a případné omezené znalosti angličtiny. Poslední dva ukazatele však byly dostupné pouze za roky 1989–1999. Od roku 1999 byly navíc k dispozici údaje o tom, zda má žák nárok na oběd zdarma nebo za sníženou cenu. Skutečnost, že některé údaje nebyly k dispozici za celé sledované období, nepředstavovala závažný problém, protože odhady přidané hodnoty učitelů nejsou citlivé na vyloučení výše zmíněných demografických charakteristik z modelů.

Datový soubor informací o žácích bylo možné spojit s druhým datovým souborem, který obsahoval informace o jejich učitelích. Informace o učitelích pocházely z „*data management systému*“, který byl v 90. letech 20. století ve sledovaném školském distriktu postupně zaváděn. Z důvodu postupného zavádění systému do škol nebyla v prvních letech sledovaného období k dispozici data o učitelích za všechny školy daného distriktu. Například údaje o učitelích některých předmětů nebyly k dispozici před školním rokem 1994, poté se jejich pokrytí zvýšilo na zhruba 60 procent a po roce 2003 se ustálilo na přibližných 85 procentech. Aby se autoři studie ujistili, že jejich odhady přidané hodnoty učitelů nejsou zkresleny chybějícími údaji, ve studii dokazovali, že jejich závěry zůstávají stejné, i když analýzu omezí jen na podsoubor s chybějícími údaji o školách a předmětech.

Před vlastní analýzou bylo nutné hrubá data o žácích a učitelích upravit, a to z několika důvodů. Zaprvé odhady přidané hodnoty učitelů vždy závisí na výsledcích žáků v předchozích testech, proto autoři studie omezili vzorek pouze na čtvrté až osmé školní ročníky, u kterých byly vždy k dispozici předchozí výsledky žáků v testech. Zadruhé, ze vzorku bylo vyřazeno 6 procent pozorování ve třídách, ve kterých bylo více než 25 procent žáků se speciálními vzdělávacími potřebami, protože v těchto třídách může vyučovat více učitelů zároveň nebo vyučování probíhá ve speciálním výukovém režimu. Autoři studie také vynechali 2 procenta pozorování, ve kterých byl žák vyučován doma, v nemocnici nebo ve škole, která slouží výhradně zdravotně postiženým žákům. Za třetí vypustili třídy s méně než 10 žáky nebo naopak s více než 50 žáky, stejně tak jako učitele spojené s více než 50 žáky (což bylo cca 0,5 procent pozorování). A v neposlední řadě, pokud učitel během jednoho školního roku vyučoval žáky ve více školách (což se vyskytovalo u 0,3 procent pozorování), použili pouze vazbu na školu, ve které byl učitel dle personálních záznamů školy uveden jako kmenový pracovník a v ostatních školách jej označili za chybějícího.

Zároveň téměř 90 procent údajů o žácích z příslušného školského distriktu bylo možné propojit s vybranými údaji ze záznamů finančních úřadů (*US tax records*) shromažďovaných ve Spojených státech v období od roku 1996 do roku 2011. Tato data obsahovala důležité informace o rodinách žáků, o příjmu domácností¹⁸, rodinném stavu¹⁹, o vlastnictví domu²⁰ a o spoření v penzijním fondu 401(k)²¹. Autoři pracovali také s věkem matky v době narození dítěte, který byl získán z údajů Správy sociálního zabezpečení o narození dítěte. Pokud nebylo možné dítě přiřadit k jeho rodičům, byly všechny charakteristiky rodičů označeny nulou. V regresních rovnicích, které zahrnovaly charakteristiky rodičů, se vždy pracovalo s „dummy“ proměnnou pro chybějící údaje o rodičích.

18 Příjem byl definován jako průměrný upravený hrubý příjem v letech 2005 až 2007 toho daňového poplatníka, který poprvé požádal o daňovou slevu na dítě. Pro roky, ve kterých rodiče nepodali daňové přiznání, jim byl přiřazen příjem 0. Příjem byl vyjádřen v dolarech v jejich kupní síle v roce 2010 (v jiných letech upravován o inflaci spotřebitelských cen).

19 Rodinný stav bylo možné poznat z toho, zda rodič, který uplatňoval nárok na slevu daně na dítě, podával společné daňové přiznání, ve kterém se uvádí i údaje o manželce/manželovi.

20 Vlastnictví domu bylo indikováno tak, že daný jedinec uplatňoval slevu na dani kvůli zaplaceným úrokům z hypotéky.

21 Spoření v penzijním fondu bylo indikováno uplatňováním slevy na dani z důvodu přispívání do penzijního spoření 401(k).

Samotná analýza byla zahájena odhadováním přidané hodnoty učitelů, u kterých byla k dispozici příslušná vstupní data. Přidaná hodnota byla odhadnuta pro každého učitele v daném školním roce na základě průměrných výsledků jeho žáků v předchozích letech. Při konstrukci odhadu byl kontrolován vliv vybraných charakteristik žáků, tedy jejich výsledky v předchozích testech a demografické proměnné, což umožnilo odlišit vliv učitele od (pozorovatelného) vlivu nenáhodného umístování žáků do jednotlivých tříd. Autoři upozorňují, že se jejich přístup k modelování přidané hodnoty s výjimkou jednoho aspektu podobal tomu, který v dané době používaly školské distrikty. Rozdíl spočíval „pouze“ v tom, že školské distrikty, na rozdíl od autorů studie, obvykle předpokládaly, že je kvalita učitele v čase neměnná, a proto při předpovídání jeho kvality přikládaly stejnou váhu výsledkům testů ve všech třídách, které učitel vyučoval. V praxi ovšem výsledky testů žáků z tříd, které učitel učil nedávno, lépe předpovídají aktuálně měřenou kvalitu učitele, což naznačuje, že kvalita učitele se v čase mění. Autoři studie proto ve své analýze zohlednili tento „posun“ v kvalitě učitele tím, že ne-parametricky odhadovali auto kovarianci skóre napříč třídami, které daný učitel vyučoval, a za pomoci regresního modelu upravovali výsledky v roce t s ohledem na průměrné výsledky v ostatních letech. Z takového modelu přidané hodnoty učitele vyplývá, že 1 směrodatná odchylka v přidané hodnotě učitele zvýší normalizované výsledky testů žáků přibližně o 0,14 směrodatné odchylky v matematice a o 0,10 směrodatné odchylky v angličtině, což je o něco více než odhady vypočítané v předchozích studiích, které posun ve schopnostech učitelů nezohledňovaly.

Následně se autoři zaměřili na nalezení odpovědi na hlavní výzkumnou otázku první studie, a to zda jsou jimi zkonstruované míry přidané hodnoty nezkreslenými prediktory kvality pedagogické práce učitele. Aby mohli formálně definovat zkreslení, použili hypotetický experiment, v němž náhodně přiřazovali žáky k učitelům, obdobně jako tomu bylo ve studii Kaneho a Staigera (2008). Nejprve vypočítali průměrný dopad náhodného přiřazení učitele žákům na jejich skóre v testu, a to pro učitele, který má na základě údajů z předchozích školních let o jednu směrodatnou odchylku vyšší přidanou hodnotu než předchozí učitel. Pokud takový bude učit žáky a oni budou následně mít o jednu směrodatnou odchylku vyšší skóre v testech, autoři udávají, že budou takové odhady přidané hodnoty považovat za „prognosticky nezkreslené“.

Pro odhad míry zkreslení odhadu přidané hodnoty učitele vyvinuli autoři studie dvě metody. Prostřednictvím první metody bylo zkreslení prognózy odhadováno na základě toho, jak byly výsledky žáků v testech ovlivněny odlišnostmi v jejich pozorovatelných charakteristikách, jestliže tyto charakteristiky nebyly použity při výpočtu přidané hodnoty učitele. Pro každého žáka vypočítali odhady výsledků standardizovaných testů na základě charakteristik rodičů (např. příjem rodiny) a odhadu přidané hodnoty učitele. V souladu s jejich základním modelem přidané hodnoty učitele, který kontroluje vliv bohatého souboru kontrolních proměnných charakterizujících předchozí výsledky žáků, třídy, školy a demografické údaje, zjistili, že zkreslení odhadů způsobené nepoužitím charakteristik rodičů je maximálně 0,3 procent při použití 95procentního intervalu spolehlivosti. Při použití podobné metodologie zjistili, že zkreslení prognózy v důsledku nepoužití skóre z předchozích testů je maximálně 2,6 procent.

Při interpretaci těchto výsledků je důležité poznamenat, že děti rodičů s vyššími příjmy mají přidanou hodnotu učitelů v průměru vyšší. To naznačuje, že v praxi dochází k sociálně selektivnímu přiřazování žáků učitelům. Autoři studie však upozorňují, že takovéto sociálně selektivní přiřazování žáků a učitelů nevede ke zkresleným odhadům přidané hodnoty učitele, a to ze dvou důvodů. Za prvé, a to je nejdůležitější, korelace mezi odhady

přidané hodnoty učitele a charakteristikami rodičů zmizí, jakmile jsou kontrolovány výsledky testů v předchozím školním roce. Za druhé, i nepodmíněná korelace mezi rodiči a odhady přidané hodnoty učitele je malá: autoři odhadují, že zvýšení příjmu rodičů o 10 000 dolarů je spojeno se zvýšením přidané hodnoty učitelů o méně než 0,0001 její směrodatné odchyly (měřeno ve směrodatné odchylyce výsledků testů žáků). Jedno z vysvětlení, proč je vliv sociálně podmíněného přiřazování žáků učitelům tak malý, je, že 85 procent variability v míře přidané hodnoty učitelů lze indikovat spíše uvnitř škol než mezi školami. Vzhledem k tomu, že většina sociálně podmíněného přiřazování žáků učitelům v praxi probíhá prostřednictvím volby školy, ale nikoliv prostřednictvím přiřazování žáků ke konkrétním učitelům v rámci školy, mohou mít rodiče jen malý prostor k tomu, aby své děti nasměrovali k učitelům s vyšší přidanou hodnotou.

Zatímco první metoda ukazuje, že zkreslení odhadů způsobené přiřazováním žáků učitelům podle určitých pozorovatelných charakteristik žáků je minimální, zkreslení způsobené jinými nepozorovatelnými charakteristikami může být stále značné. Aby autoři získali přesnější odhad míry zkreslení odhadu přidané hodnoty učitelů, který zohledňuje i nepozorovatelné charakteristiky, vyvinuli kvazi-experimentální analogii ideálního experimentu, kterým je náhodné přidělování žáků učitelům. Použitá kvazi-experimentální metoda využívá skutečnosti, že dochází k fluktuaci učitelů na úrovni školních tříd. Metodologický design je založen na předpokladu, že učitel třídy ve 4. ročníku, u kterého byla naměřena vysoká přidaná hodnota, se například na začátku roku 1995 přestěhuje na jinou školu. V důsledku této personální změny budou mít žáci dané třídy ve 4. ročníku ve škole „A“ v roce 1995 v průměru učitele s nižší přidanou hodnotou než předchozí kohorta žáků. Pokud by odhady přidané hodnoty učitele měly schopnost predikovat výsledky žáků v testech, dalo by se očekávat, že výsledky testů žáků sledované třídy ve škole „A“ v roce 1995 budou v průměru nižší než v předchozí kohortě.

Studiem situací s příchody a odchody učitelů autoři zjistili, že průměrné výsledky žáků v testech napříč jednotlivými kohortami se opravdu prudce mění v souladu s předpokladem, že výsledky žáků ovlivňují příchody a odchody učitelů s velmi vysokou/nízkou přidanou hodnotou ze tříd. Velikost zkreslení takové predikce autoři odhadovali prostřednictvím porovnání změn v průměrných výsledcích testů v po sobě jdoucích kohortách žáků v rámci školy s ohledem na změny průměrné přidané hodnoty pedagogického sboru. Autoři dospěli k závěru, že se předpokládané změny průměrných hodnot skóre v testech přesně shodují s reálně pozorovanými změnami v učitelském sboru. Bodový odhad zkreslení prognózy v jimi preferované specifikaci činí 2,6 procent a není statisticky odlišitelný od 0. Horní hranice pro velikost zkreslení činí 9,1 procent (při 95procentním intervalu spolehlivosti). Použitý kvazi-experimentální design přitom vycházel z předpokladu, že vysokofrekvenční fluktuace učitelů mezi jednotlivými třídami v rámci školy není korelována s charakteristikami žáků a charakteristikami školy. Tento předpoklad je věrohodný, neboť je nepravděpodobné, že by rodiče okamžitě převedli své dítě na jinou školu jen kvůli odchodu nebo příchodu jednoho učitele. Kromě toho ukazují, že změny v průměrné kvalitě učitelů v daném předmětu (např. matematika) nejsou korelovány ani s předchozími výsledky v tomto předmětu, ani se současnými výsledky v jiných předmětech (např. angličtina), což podporuje platnost výzkumného designu.

Autoři zároveň zkoumali, které z kontrolních proměnných v jejich základním modelu přidané hodnoty učitelů jsou nejdůležitější při měření vlivu učitele na výsledky žáka. Počítali proto zkreslení míry odhadů výsledků žáků pomocí výše uvedeného kvazi-experimentálního modelu pro několik běžně používaných měř přidané hodnoty učitelů. Bylo zjištěno, že pokud je kontrolován jen vliv předchozích výsledků žáka v testech, odhad zkreslení

prognózy dosahuje cca 5 procent, což se významně neliší od nuly. Naproti tomu modely, které záměrně vynechávají kontrolu vlivu skóre žáka v předchozích testech, vytváří zkreslení prognózy přesahující 40 procent. Většina efektu nenáhodného přiřazování žáků učitelům, které je relevantní pro prognózu budoucích výsledků v testech, je zachycena předchozími výsledky v testech. Tento výsledek je příznivý pro použití modelů přidané hodnoty učitelů v praxi, protože prakticky všechny modely s přidanou hodnotou používané v praxi kontrolují vliv skóre žáků v předchozích testech.

2.1.2 Diskuze výsledků v kontextu dosavadní odborné literatury

Kvazi-experimentální metoda poskytuje jednoduchý a levný nástroj pro hodnocení zkreslení v různých podmínkách. Například Kane, Staiger a Bacher-Hicks (2014) aplikují tuto metodu na datech z Los Angeles z *Unified School District*. Uvádějí, že prognózy vlivu přidané hodnoty učitele, při které je kontrolován vliv výsledků předcházejících testů absolvovaných žáky, rovněž nevykazují žádné zkreslení, přestože rozptýl přidané hodnoty učitelů je ve zkoumaném školském distriktu mnohem větší než ve školním distriktu, který byl použit v této studii. Obecněji řečeno, Chetty a kol. (2014a) vyvinuté metody by mohly být použity k posouzení přesnosti metrik hodnocení pracovníků i v různých jiných profesích mimo samotné učitelství.

Výsledky této studie navíc dávají do souladu zjištění experimentálních studií Kanea a Staigera (2008) a Kane a kol. (2013) s výsledky experimentální studie prováděné Rothsteinem (2010). Chetty a kol. (2014a) potvrdili Rothsteinovo zjištění, že existuje malé ale statisticky významné nenáhodné přiřazování žáků učitelům, které spočívá v tom, že přírůstky testového skóre žáků v jejich minulých testech souvisí s jejich přiřazením jednotlivým učitelům. Ukazují ovšem, že tento konkrétní zdroj nenáhodné selekce žáků vytváří jen minimální zkreslení prognóz. Rothstein ovšem varuje, že přiřazování učitelů žákům na základě nepozorovatelných veličin může potenciálně způsobit značné zkreslení. Chetty a kol. (2014a) posuzují míru zkreslení odhadu přidané hodnoty učitele způsobenou nepozorovatelnými veličinami pomocí kvazi-experimentální obdoby experimentu Kanea a Staigera (2008). Stejně jako Kane a Staiger ovšem nenašli žádné důkazy o zkreslení odhadu v důsledku vlivu nepozorovatelných faktorů. Proto dospěli k závěru, že přidaná hodnota učitele, při jejímž odhadu je kontrolován vliv předchozích výsledků žáků v testech, vykazuje jen malé zkreslení navzdory nenáhodnému přiřazování žáků učitelům podle jejich předchozích testových výsledků.

Chetty a kol. (2014a) udávají, že i když předchozí výzkumy zkreslení v modelech přidané hodnoty docházejí k rozporuplným závěrům (např. Kane a Staiger 2008, Rothstein 2010, Koedel a Betts 2011, Kinsler 2012, Goldhaber a Chaplin 2012, Kane a kol. 2013), jejich zjištění pomáhají tyto rozpory uvést do souladu a ukazují, že v literatuře de facto panuje shoda v hlavních empirických otázkách. Rothstein (2009, 2010) inicioval výzkum zkreslení v modelech přidané hodnoty učitele tím, že ve svých analýzách dospěl ke dvěma základním výsledkům. Za prvé ukázal, že existuje významné nenáhodné seskupování žáků do tříd na základě jejich výsledků v testech v předminulém školním roce, ovlivněné i výsledky jejich testů z minulého roku (Rothstein 2010). Za druhé prokázal, že toto nenáhodné přiřazování žáků učitelům vytváří jen minimální zkreslení v odhadech přidané hodnoty učitele: kontrola skóre žáků z předminulého roku nemá významný vliv na odhad přidané hodnoty učitele (Rothstein 2010). Tyto výsledky se opakují v těchto datech obdobně jako ve studii Kaneho a Staigera (2008). Lite-

ratura se tedy shoduje na tom, že opatření, které využívají měření přidané hodnoty učitele, netrpí podstatným zkreslením, pokud kontrolují vliv výsledků žáků v předchozích testech.

Rothstein (2010) zdůrazňuje, že jeho zjištění vzbuzují obavy ohledně možnosti zkreslení v důsledku nenáhodného přiřazování žáků a učitelů prováděného na základě nesledovaných proměnných charakterizujících žáky či učitele. Kane a Staiger (2008) a Kane a kol. (2013) se snažili analyzovat data, aby prozkoumali možné odpovědi na tuto obavu. Přitom uvádějí odhady zkreslení z experimentálních studií, v nichž byli žáci přiřazeni k učitelům náhodně. Jejich odhady naznačují, že vliv nenáhodného přiřazování žáků učitelů na základě nepozorovatelných charakteristik je malý, ale jejich konfidenční intervaly spočítané na 95procentní hladině spolehlivosti byly velké – z důvodu malé velikosti výběrového souboru byly v některých případech až padesátiprocentní.

Kvazi-experimentální odhady Chettyho a kol. (2014a) ukazují, že míra zkreslení způsobená nenáhodným přiřazováním žáků učitelům na základě jejich nepozorovatelných charakteristik se ukazuje jako zanedbatelná, pokud se pracuje s mnohem větším reprezentativnějším vzorkem. Nedávná replikace této metodiky realizovaná Kanem a kol. (2014), kteří použili data ze školského distriktu v Los Angeles, rovněž dochází ke stejnému závěru. Z toho vyplývá, že studie, které přímo testovaly vliv nenáhodného přiřazování žáků učitelům na základě nepozorovatelných charakteristik, se shodují v tom, že odhady přidané hodnoty učitelů nejsou těmito faktory významně zkresleny. V budoucím výzkumu by mohlo být zajímavé prozkoumat, proč nenáhodné přiřazování žáků učitelům na základě skóre v předchozích testech, původně zdokumentované Rothsteinem (2010), nakonec nevede k významnému zkreslení prognózy přidané hodnoty učitelů provedené modelovými odhady. Zjištění uvedená v tomto článku a v související literatuře však postačují k závěru, že standardní odhady přidané hodnoty učitelů poskytují nezkreslené odhady vlivu učitelů na výsledky žáků v testech.

2.1.3 Závěr

Chetty a kol. (2014a) ukazují, že odhady přidané hodnoty učitelů založené na zkoumání výsledků testů jejich žáků jsou užitečnými prediktory kvality učitelů, a že lze tyto predikce potenciálně zlepšit nejméně ve dvou dimenzích. Zaprvé, začleněním dalších ukazatelů kvality učitelů (např. hodnocení učitele ředitelem) nebo informací o dalších charakteristikách učitelů může přinést lepší výsledky. Je možné použít kvazi-experimentální metodiku vyvinutou v práci Chetty a kol. (2014a) k ověření validity takových ukazatelů kvality učitelů. Například když učitel s dobrým hodnocením od ředitele odejde na jinou na školu, zlepší se výsledky žáků v nové škole jeho působení? Zadruhé, by bylo cenné vyvinout lepší ukazatele dopadů kvality učitelů, které by měly lepší vypovídací schopnost než používané ukazatele průměrného skóre jejich žáků v testech, které byly v tomto textu analyzovány. Bylo by možné vyvinout nástroje pro měření přidané hodnoty učitelů, které by pracovaly s různými percentily rozdělení výsledků testů žáků a s výsledky testů pro konkrétní demografické podskupiny žáků. Jsou například někteří učitelé lepší při vzdělávání chlapců než při vzdělávání dívek nebo při vzdělávání žáků s celkově dobrými výsledky než při vzdělávání žáků s celkově špatnými výsledky v testech?

V tomto článku bylo zjištěno, že měření přidané hodnoty nám může pomoci určit, kteří učitelé mají největší schopnost zvýšit průměrné výsledky žáků v testech. Z tohoto výsledku ovšem nelze automaticky vyvodit závěr,

že učitel s vysokou změřenou přidanou hodnotou je „dobrý učitel“, protože výsledky žáků ve standardizovaných srovnávacích testech nejsou konečným výsledkem, který nás nejvíce zajímá. Tím je, zda vysoká přidaná hodnota učitele změřená tímto způsobem také zlepšuje dlouhodobé výsledky žáků, nebo jsou tito učitelé prostě jen lepší ve výuce, která má za cíl připravit žáky na úspěšné absolvování testů.

2.2 Zlepšují učitelé s vysokou přidanou hodnotou dlouhodobé výsledky žáků²²

Druhá studie autorů Chettyho, Friedmana, and Rockoffa (2014b) se zaměřuje na analýzu dlouhodobého vlivu učitelů, u nichž byla naměřena vyšší přidaná hodnota, na jeho žáky. To bylo ve studii konkrétně specifikováno jako hledání odpovědi na otázku, zda ti učitelé, kteří zvyšují výsledky žáků v testech, také zároveň zlepšují výsledky svých žáků v dospělosti. Metodologický design studie byl založen na propojení informací ze dvou různých zdrojů. Jednalo o propojení informací získaných ze školských úřadů (konkrétně o datový soubor o žácích a učitelích třetích až osmých ročníků velkého městského školského distriktu, který byl shromážděn v letech 1989–2009) s vybranými údaji z daňových záznamů amerického finančního úřadu shromážděnými v letech 1996–2011. Chettymu a kol. (2014b) se podařilo propojit zhruba 90 procent záznamů o žácích získaných během jejich školní docházky s údaji z daňových přiznání bývalých žáků, což umožnilo sledovat přibližně jeden milion jedinců od základní školy až po ranou dospělost. Z daňových přiznání bylo zjištěno, jaké byly po absolvování školy jejich výdělky, zda chodili na vysokou školu a rovněž některé další demografické údaje (např. porodnost).

2.2.1 Data, metodologie, shrnutí výsledků

Jak bylo uvedeno výše, Chetty a kol. (2014b) všechny informace čerpali ze dvou různých zdrojů: (1) z administrativní evidence školských distriktů a (2) ze záznamů finančního úřadu získaných z formulářů, které občané Spojených států vyplňují při podávání daňového přiznání k federální dani z příjmu. Datová sada použitá v tomto článku byla totožná s datovou sadou použitou v prvním článku s tím rozdílem, že byla použita pouze data o podskupině bývalých žáků, kteří byli již dostatečně staří na to, aby bylo možné nejpozději v roce 2011 pozorovat výsledky jejich pracovní a životní dráhy v jejich rané dospělosti.

Informace o žácích obsahovaly historii jejich zápisů do jednotlivých škol a ročníků, skóre testů, které během své školní docházky absolvovali, a zároveň informace o učitelích, kteří jim byli přiřazeni. Všechny zmíněné údaje pocházely ze školních roků počínajících školním rokem 1988/1989 a končících školním rokem 2008/2009. Celkově byla shromážděna data o přibližně 2,5 milionu žácích zapsaných ve sledovaných letech ve třetím až osmém ročníku základní školy. Dohromady byly shromážděny skóre žáků z 18 milionů testů, které měřily jimi dosaženou úroveň v anglickém jazyce a matematice. Výsledky všech testů byly normalizovány na tzv. z-skóre, tedy tak, aby měl průměrný výsledek u každého žáka v testu v každém roce a v každém testovaném ročníku hodnotu nula. Přitom kladné hodnoty znamenaly lepší než průměrné a záporné hodnoty horší než průměrné výsledky a zá-

²² Tato kapitola je podrobným výtahem z textu: Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9), 2633–2679. Z tohoto textu je přebírán i odkazový aparát.

roven „vzdálenost od nuly“ byla měřena v násobcích směrodatné odchylky, což zajišťovalo, že rozptyl výsledků v rámci ročníku ve zkoumaném školském distriktu byl srovnatelný s rozptylem pozorovaným v celostátním vzorku výsledků testů v příslušném ročníku. To umožňovalo porovnávat výsledky s odhady založenými na analýze jiných výběrových vzorků testovaných žáků.

Datový soubor o žácích obsahoval mimo jiné také informace o jejich etnickém původu, pohlaví, věku a rovněž informace o tom, zda patřili mezi žáky se speciálními vzdělávacími potřebami či mezi žáky s omezenou znalostí angličtiny a to separátně pro všechny zkoumané školní roky. Od školního roku 1998/1999 byly navíc k dispozici informace o tom, který žák měl nárok na oběd zdarma nebo oběd za sníženou cenu. Ke každému z žáků, o nichž byly shromážděny výše zmíněné informace, byly zároveň k dispozici určité informace o jejich učitelích, i když v prvních sledovaných letech chyběly údaje z některých škol, zejména údaje o učitelích působících na *middle schools*²³ byly dost omezené. V průběhu analýzy proto byly provedeny statistické testy, jejichž výsledky ukázaly, že celkové výsledky analýzy nebyly ovlivněny omezeným množstvím informací či chybějícími informacemi o některých učitelích z některých škol.

Počáteční soubor dat o žácích a jejich učitelích byl obdobně jako v případě první studie následně o něco zmenšen. Důvodem bylo, že odhady přidané hodnoty učitelů vždy závisí na testových výsledcích jejich žáků. Z tohoto důvodu byl vzorek žáků a jejich učitelů omezen jen na žáky čtvrtých až osmých ročníků, u nichž byly vždy k dispozici výsledky předchozích testů. Bylo také vyloučeno přibližně 6 procent výsledků testů, které byly získány ve třídách, kde bylo více než 25 procent žáků se speciálními vzdělávacími potřebami, protože v takových třídách mohou být aplikovány speciální vzdělávací programy nebo mohou mít více učitelů ve třídě. Podobně byla vyloučena přibližně 2 procenta testových výsledků od žáků, kteří jsou sice zapsáni ve škole ve zkoumaném školském distriktu, ale ve skutečnosti jsou vzdělávány převážně formou domácího vzdělávání či vzdělávány v nemocnici, nebo jsou zapsáni ve škole sloužící výhradně handicapovaným žákům. Z analýzy byla vyloučena také všechna pozorování ze tříd, ve kterých bylo udáváno méně než 10 žáků nebo více než 50 žáků a také všechna pozorování od učitelů, u kterých bylo udáváno, že vyučovali více než 200 žáků stejného ročníku, neboť se pravděpodobně jedná o chybně zaznamenaný údaj (tyto záznamy tvořily 0,5 procenta případů). Pokud byl učitel propojen se žáky ve více školách (0,3 procent případů) pracovalo se pouze s daty ze školy, ve které bylo učitelovo hlavní zaměstnání. Vzhledem k tomu, že údaje o pracovní a životní dráze v rané dospělosti pocházející z daňových přiznání jedinců ve věku 20 let a více, byla data ze škol omezena jen na žáky, kteří do roku 2011 dovršili nejméně 20 let.

Jak již bylo zmíněno, data o pracovní a životní dráze žáků po absolvování školy byla získána z daňových přiznání k federální dani z příjmu za období 1996/2011. Tyto údaje byly propojeny se záznamy ze školského distriktu prostřednictvím algoritmu založeného na standardních identifikátorech (datum narození, stát narození, pohlaví a jméno), která se vyplňují do přílohy daňového přiznání. Po propojení byly tyto identifikátory ze souboru odstraněny a data byla anonymizována. Celkově se podařilo propojit data z daňových přiznání s údaji 87,4 procenta žáků.

Výsledky bývalých žáků v dospělosti byly definovány následujícím způsobem. Údaje o výdělcích jedinců pocházejí z daňových formulářů W-2. Tyto údaje jsou dostupné jak pro jedince, kteří podávají daňové přiznání, tak

23 Odpovídá druhému stupni základní školy.

i pro jedince, kteří daňové přiznání sami nepodávají. Roční výdělky vyšší než 100 000 USD byly kódovány jako 100 000 US, aby se snížil vliv odlehlých hodnot na výsledky analýz (ve věku 28 let takových výdělků dosahovalo 1,3 procenta jedinců). Všechny výdělky byly zároveň přepočítány na USD v roce 2010 prostřednictvím úpravy o inflaci měřenou vývojem spotřebitelských cen. Celkem 33,1 procent jedinců mělo ve věku 28 let nulové výdělky. Kromě individuálních výdělků byly definovány i celkové příjmy. Ty se vypočítaly jako souhrn všech individuálních výdělků členů domácnosti a k nim se navíc připočetly příjmy ze samostatné výdělečné činnosti členů domácnosti. Celkem 29,6 procent jedinců mělo takto definované nulové celkové příjmy.

Docházka jedinců na vysokou školu byla indikována tak, že všechny vysoké školy a univerzity, jakož i odborné školy a další postsekundární instituce způsobilé pro získání federální dotace pro své studenty jsou povinny podávat finančnímu úřadu daňové formuláře 1098-T, které u každého studenta udávají, že zaplatil školné nebo naopak získal stipendium. Nezávisle na tom, zda fyzická osoba podává či nepodává daňové přiznání, lze proto z údajů finančního úřadu zjistit o všech jedincích kompletní záznamy o jejich docházce na vysokou školu. Tyto údaje z vysokých škol navíc umožňují vytvořit index kvality vysoké školy podle toho, jakou konkrétní vysokou školu bývalí žáci ve věku 20 let navštěvovali. Pro všechny, kteří navštěvovali vysokou školu nebo univerzitu, ale také pro ty, kteří nenavštěvovali žádnou vysokou školu, byly z údajů finančního úřadu vypočítány průměrné výdělky ve věku 31 let (v letech 2010 a 2011). Mezi žáky ze zkoumaného školského distriktu, kteří byli předmětem této studie, dosáhla průměrná hodnota indexu výdělků 44 048 USD u jedinců, kteří byli ve 20 letech zapsáni ve čtyřletých vysokoškolských programech a 30 946 USD v případě jedinců zapsaných ve dvouletých vysokoškolských programech. Žáci, kteří ve 20 letech nebyli na žádné vysoké škole, dosáhli průměrné úrovně výdělků 17 920 USD.

Údaje z finančního úřadu navíc umožnily zjistit adresu každého jedince v databázi a následně zkonstruovat indikátor socioekonomického statusu čtvrti, v níž bydlí, a to pomocí údajů o procentu vysokoškolsky vzdělaných obyvatel čtvrti dle údajů ze sčítání lidu v roce 2000. Kromě toho údaje daňového úřadu poskytly informace o tom, zda mají jedinci penzijní spoření na tzv. účet 401(k). Data z finančního úřadu ovšem přinesla i některé další demografické a socioekonomické informace. Umožnila například zjistit, zda žena v databázi finančního úřadu byla/nebyla matkou dítěte už ve věku 13 až 19 let. V základním vzorku bylo 15,8 procenta takových žen. Umožnila také zjistit, některé charakteristiky rodičů zkoumaných jedinců, které bylo následně možno využít jako kontrolní proměnné. Údaje o rodičích bylo možno identifikovat u 94,8 procent zkoumaných jedinců. Šlo především o příjem rodičů, rodinný stav rodičů, vlastnictví domu, penzijní spoření 401(k) a věk matky při narození dítěte.

Do vlastní analýzy nakonec vstupovalo celkem 6,8 miliónů pozorování, které pocházelo od celkem 1,1 milionu žáků. U každého žáka, jehož životní a pracovní dráha v mladé dospělosti byla následně zkoumána, bylo k dispozici v průměru 6,25 pozorování (jedno pozorování je definováno jako výsledky předmětového testu v konkrétním školním roce).

Při vlastní analýze byla nejprve provedena deskriptivní analýza, která poskytla průměry a směrodatné odchylky u všech proměnných. Zároveň deskriptivní analýza ukázala, že v souboru je celkem 1,6 milionu pozorování (výsledků předmětových testů v konkrétním ročníku), ke kterým je možné zároveň přiřadit informace o přidané hodnotě jejich učitele v době konání testu a jejich výdělcích ve věku 25 let. Jedinců s plnou informací o výsledcích

testů a přidané hodnotě učitele, u kterých jsou zároveň známy výdělky ve věku 28 let, je 750 tisíc, ale jen u 220 tisících takových jedinců jsou známy výdělky ve věku 30 let. Ve druhém kroku byla provedena regresní analýza (klasická metoda OLS regrese), která hledala statistické souvislosti mezi výsledky testů z dob školní docházky zkoumaných jedinců a jejich úspěšností v mladém dospělém věku. Výsledky testů z matematiky i angličtiny jsou vysoce pozitivně asociovány s vyšší výdělků dosažených v mladém dospělém věku, s docházkou na vysokou školu a i kvalitou čtvrti, ve které jedinci bydlí. Zároveň analýza ukázala negativní asociaci s pravděpodobností, že bývalá žákyně porodí dítě před dvacátým rokem věku. V průměru je zvýšení výsledku standardizovaného testu o jednu směrodatnou odchylku nad průměrnou hodnotu spojeno se zvýšením o výdělek o 7 700 USD (tj. o 36 procent oproti průměru) ve věku 28 let. Pokud se ovšem do modelů zahrne vliv kontrolních proměnných, zlepšení skóre v testu o jednu standardní odchylku v aktuálním testu je spojeno s průměrným zvýšením výdělku o 2 600 USD (tj. o 12 procent). Vztah mezi skóre ve školních testech a budoucími výsledky v mladém dospělém věku má tvar písmene S. Velký vliv mají skóre z testů předcházejících období měření, vliv dalších kovariátů je téměř dokonale lineární. Kauzální dopady přidané hodnoty učitele na výdělky a další výsledky jejich bývalých žáků jsou úměrné tomu, co by se dalo předpovědět na základě těchto statistických vztahů.

V další fázi analýzy autoři studie využili kvazi-experimentální design, aby se vyrovnali s potenciální námitkou, že existují nepozorované faktory ovlivňující dlouhodobé výsledky žáků, které korelují s kvalitou učitele, jež je podmíněna pozorovatelnými veličinami. V této části odhadovali dlouhodobé dopady učitelů pomocí designu, který využívá skutečnosti, že učitelé se někdy ve třídách mění. Tak se může stát, že učitel s vysokou přidanou hodnotou učitele je nahrazen učitelem s nízkou přidanou hodnotou nebo naopak. Tato analýza prokázala, že změny v kvalitě pedagogického sboru mají významný dopad změny v míře docházky na vysokou školu ve věku 20 let napříč po sobě jdoucími kohortami žáků a stejně tak mají statisticky významný dopad na průměrnou kvalitu vysoké školy, které žáci navštěvují. Ukázalo se také, že charakteristiky žáka a rodičů nekorelují se změnami v kvalitě pedagogických pracovníků. Výsledky předchozích testů žáků a skóre v jiném předmětu také nekoreluje se změnami v kvalitě pedagogických pracovníků v daném předmětu. Změny v přidané hodnotě učitelů mají silný vliv na pravděpodobnost studia na vysoké škole. Změna pravděpodobnosti, že bude bývalý žák studovat na vysoké škole, byla pozorována přesně v roce změny učitele, ale ne v předchozích letech nebo následujících letech.

Nakonec se autoři studie zaměřili na analýzu vlivu vybraných demografických faktorů, konkrétních předmětů a ročníků, ve kterých proběhlo standardizované testování žáků. Aby se maximalizovala přesnost odhadů, byly odhadovány dopady těchto faktorů na kvalitu vysoké školy, ve které byl bývalý žák zapsán ve věku 20 let, protože bylo v datovém souboru více případů s těmito údaji než případů s informacemi o výši výdělku ve vyšším věku. Výsledky analýzy prokázaly, že takto měřené dlouhodobé dopady přidané hodnoty učitelů jsou o něco větší u žen než u mužů, pokud je kontrolován vliv dalších individuálních charakteristik žáka a charakteristik jeho rodičů. Analýzy také prokázaly, že dopady jsou větší u žáků pocházejících z rodin s vyššími příjmy než u žáků z rodin s nižšími příjmy. Například zvýšení přidané hodnoty učitelů o jednu standardní odchylku vedlo u žáků z rodin s podprůměrnými příjmy k takovému zvýšení kvality školy, které odpovídá odhadovanému vyššímu příjmu jejich absolventů o 190 USD ročně, zatímco žáků z rodin s nadprůměrnými příjmy vedlo ke zvýšení kvality školy, které odpovídalo odhadovanému vyššímu příjmu jejich absolventů o 380 USD ročně. V relativním vyjádření to znamená nárůst odhadovaných příjmů o 0,80 procenta pro žáky z rodin s podprůměrnými příjmy oproti 1,25

procenta pro žáky z rodin s nadprůměrnými příjmy. Větší absolutní dlouhodobý dopad na žáky z rodin s nadprůměrnými příjmy může být teoreticky způsoben buď tím, že zvýšení přidané hodnoty učitele více zvýší skóre takových žáků v testech, nebo se skóre zvýší podobně jako u žáků z rodin s podprůměrnými příjmy, ale jeho dlouhodobé pozitivní důsledky na vzdělávací (a následně i pracovní) dráhu žáka budou větší. Analýza prokázala, že druhé z možných vysvětlení je podstatně pravděpodobnější: rozdíly mezi žáky z nadprůměrně a podprůměrně vydělávajících rodin je především v dlouhodobých dopadech. Obdobně se také prokázalo, že dlouhodobý vliv přidané hodnoty učitele je větší u žáka z domácnosti, která patří k většinové populaci, než u žáka z domácnosti patřící k menšinám. To naznačuje, že zvýšení přidané hodnoty učitelů zlepšuje výsledky žáků v rané dospělosti ve vzájemné interakci s působením dalších socio-ekonomických i kulturních zdrojů na straně rodin, které jsou vyšší u rodin s vyššími příjmy a rodin patřících k etnické a kulturní majoritě ve společnosti.

Autoři studie si rovněž položili otázku, jestli působení změn v přidané hodnotě učitele na budoucí životní a pracovní dráhu bývalých žáků nějakým způsobem souvisí s předmětem, který učitelé vyučují. Analýzy ukázaly, že zvýšení přidané hodnoty učitelů angličtiny má větší dlouhodobý pozitivní dopad než zvýšení přidané hodnoty učitelů matematiky. Analýza prokázala, že zvýšení o přidané hodnoty učitele angličtiny o jednu směrodatnou odchylku má přibližně 2,3 krát větší dopad na kvalitu vysoké školy, do které je ve 20 letech zapsán bývalý žák, než analogicky velké zvýšení přidané hodnoty učitele matematiky o jednu směrodatnou odchylku. Tato souvislost byla prokázána jak na 1. prvním stupni, tak na 2. stupni základní školy. Platí ovšem zároveň, že pro učitele je obecně mnohem těžší zlepšit výsledky žáků v angličtině než v matematice (viz např. Hanushek, Rivkin 2010; Kane a kol., 2013).

Nakonec se autoři studie zaměřili na zkoumání toho, zda dlouhodobý vliv změny v přidané hodnotě učitele na životní a pracovní dráhu jejich žáků nějakým způsobem souvisí s ročníkem, ve kterém žák byl, když byla změna přidané hodnoty učitele zaznamenána. Analýza prokázala, že dlouhodobý dopad učitelů je velký a statisticky významný, ať už k němu dojde v jakémkoliv ročníku. V datech nebyly nalezeny žádné systematické trendy, které by dokládaly odlišnosti mezi ročníky. To, že dlouhodobý vliv změn přidané hodnoty učitelů je víceméně stejný v jakémkoliv ze zkoumaných ročníků dokazuje, že zlepšení kvality vzdělávání v jakémkoliv období, od raného dětství až po starší školní věk, má velkou dlouhodobou návratnost a pozitivně ovlivňuje životní a pracovní dráhy žáků.

2.2.2 Relevance výsledků pro vzdělávací politiku

Chetty a kol. (2014b) v závěrečné části studie soustředili svoji pozornost na potenciální důsledky svých zjištění o dlouhodobých dopadech přidané hodnoty učitelů na životní a pracovní dráhy žáků a jejich potenciální relevanci pro vzdělávací politiku. Využili data o přidané hodnotě učitelů a data o příjmech jejich bývalých žáků v mladém dospělém věku, aby se pokusili odhadnout, jaké dlouhodobé ekonomické dopady mohou mít různé vzdělávací politiky. Při modelové analýze ekonomických dopadů vycházeli z několika předpokladů. Předně předpokládali, že procentuální dopad zlepšení přidané hodnoty učitele o jednu směrodatnou odchylku na výdělek pozorovaný u jeho bývalých žáků ve věku 28 let je během životního cyklu konstantní (a je na úrovni 1,34 procenta výdělku). V modelech ignorovali možnost, že by se mohly snížit rozdíly ve výdělcích, pokud by všechny děti dosáhly lepší-

ho vzdělání. Celoživotní příjmy průměrného amerického dítěte ve věku 12 let (což je průměrný věk testovaných žáků v souboru) odhadli ve shodě Kruegerem (1999) na 522 000 USD (v cenách roku 2010). Nakonec předpokládali, že přidaná hodnota učitele má mezi učiteli normální rozdělení.

Analytický design spočíval v myšlence, že vzdělávací politika může ovlivnit kvalitu učitelů, tedy jejich přidanou hodnotu učitele, a tím mít dlouhodobý vliv na výdělky bývalých žáků v jejich dospělém věku. Aby mohli autoři provést odhad celospolečenské hodnoty důsledků zvyšování přidané hodnoty učitele, vyšli z návrhů Hanusheka (2009, 2011) nahradit ve školách učitele, jejichž hodnocení přidané hodnoty učitele je ve spodních 5 procentech distribuce, učiteli průměrné kvality. Při vlastní analýze autoři odhadovali dopad takového opatření vzdělávací politiky. Pro zjednodušení vypočítali tyto potenciální dopady na budoucí příjmy bývalých žáků jen pro učitele základních škol, kteří vyučují jednu třídu za rok. Výsledky analýzy ukázaly, že nahrazení 5 procent učitelů s nejnižší naměřenou přidanou hodnotou učiteli s průměrnou přidanou hodnotou by zvýšilo budoucí celoživotní výdělky jednoho žáků v průměru o 14 500 USD (tedy z 522 000 USD na 536 500 USD, v cenách roku 2010). Protože učitel neučí jediné dítě, ale celou třídu, která má v průměru 28,2 žáků, by průměrné celkové zvýšení celoživotních příjmů žáků ve třídě dosáhlo cca 407 000 USD. Potenciální dlouhodobé celospolečenské zisky ze zvýšení kvality učitelů jsou tedy velké. Pokud by proces nahrazování učitelů s nejnižší přidanou hodnotou pokračoval po nějakou dobu, průměrná přidaná hodnota učitele by se sice zvyšovala, ale přidaná hodnota 5 procent učitelů s nejnižší přidanou hodnotou by se zvyšovala rychleji než průměr. Analýza se pokusila odhadnout, jaký by byl realistický „dlouhodobý výnos“ z takové politiky, pokud by byla provozovaná dlouhodoběji. Autoři dospěli k závěru, že částka by byla přibližně 250 000 USD, pokud by byli učitelé propouštěni na základě přesné znalosti jejich skutečné přidané hodnoty. Ve skutečnosti ovšem nikdo nezná úplně přesně přidanou hodnotu učitele, ale tento ukazatel je s jistotou mírou chyby odhadován na základě výsledků žáků ve standardizovaných testech. Jestliže se vezme v potaz chyba měření, která vede ke skutečnosti, že by byli propuštěni i někteří učitelé, jejichž přidaná hodnota nepatří mezi 5 procent nejnižších, a naopak nebyli propuštěni někteří učitelé, jejichž přidaná hodnota učitele mezi 5 procent nejhorších patří, autoři studie odhadli realistikou částkou beroucí v potaz všechna omezení praktického provádění takové politiky na 184 000 USD.

V praxi by ovšem ani samotné nahrazení učitelů s nejnižší přidanou hodnotou průměrnými učiteli nebylo bez dodatečných nákladů. Ty mohou zahrnovat například náklady spojené s vyšší mírou docházky na vysokou školu a s tím spojenými výdaji na fungování vysokých škol. Jiný typ dodatečného nákladu na realizaci takové politiky by spočíval ve skutečnosti, že učitelská profese by se stala rizikovější (5 procent učitelů s nejnižší přidanou hodnotou by pravidelně přicházelo o práci), takže zvýšení tohoto rizika spojeného s prací učitele, kterému by v důsledku takového systému hodnocení všichni učitelé čelili, by bylo třeba učitelům nějakým způsobem finančně kompenzovat. Rothstein (2012) odhadnul tyto náklady pomocí strukturálního modelu trhu práce pro učitele a došel k závěru, že vzdělávací politika, která systematicky propouští učitele, pokud jejich odhadovaná přidaná hodnota po třech letech poklesla pod pátý percentil, by vyžadovala průměrné zvýšení platu 1,4 procenta u všech učitelů. Ve vzorku, který byl použit v současné studii, byl plat průměrného učitele přibližně 50 000 USD, což znamená, že by roční plat učitele musel být zvýšen o přibližně 700 USD, aby jim byly kompenzovány dodatečné náklady na zvýšené riziko propuštění. Politika systematického propouštění 5 procent učitelů s nejnižší přidanou hodnotou by na základě výpočtů generovala budoucí zisky žáků v jejich dospělosti ve výši 184 000 USD na každého propuštěného učitele, tj. v průměru 9 250 USD v přepočtu na všechny učitele (protože pouze 1 z 20 učitelů by

byl ve skutečnosti propuštěn). Proto odhad zisků generovaných takovou politikou byl více než desetkrát vyšší než náklady se zavedením a aplikací této politiky spojené.

Alternativní možnosti, jak prostřednictvím vzdělávací politiky ovlivnit přidanou hodnotu učitelů, je snaha udržet v profesi učitele s vysokou přidanou hodnotou prostřednictvím vyplácení speciálních bonusů. Autoři studie použili simulace metodou Monte Carlo, aby odhadli, že udržení učitele na devadesátém pátém percentilu odhadovaného rozdělení přidané hodnoty učitele o jeden další rok v učitelské profesi, by vedlo ke zvýšení budoucích výdělků jejich žáků v průměru o 212 000 USD v součtu pro všechny žáky ve třídě. Podle dostupných údajů se zhruba 9 procent učitelů, kteří působili po tři roky ve školství, do své práce čtvrtý rok nevrátí. Míra, s jakou učitelé ve zkoumaném vzorku opouštějí profesi, nesouvisí s přidanou hodnotou učitele, což je v souladu se zjištěními Boyda a kol. (2008). Clotfelter a kol. (2008) odhadli, že výplata bonusu 1 800 USD k ročnímu platu snížila v Severní Karolině míra odchodu z profese o 17 procent. Na základě těchto odhadů bylo autory současné studie odhadnuto, že vyplacení jednorázové prémie ve výši 1 800 USD učitelům s vysokou přidanou hodnotou by zvýšilo míru jejich udržení v profesi o 1,5 procentního bodu v následujícím roce a ve svém důsledku by to vygenerovalo průměrné zvýšení příjmů jejich žáků o 3 180 USD. Očekávaný přínos vyplývající z nabízení bonusu jen vynikajícím učitelům s nejvyšší přidanou hodnotou (devadesátý pátý percentil) je jenom o něco větší než náklady s tím spojené. Efektivita opatření je totiž významně snižována tím, že by musely na každého učitele s vysokou přidanou hodnotou, kterého díky aplikaci tohoto opatření ve školství udrží, platit bonusy přibližně 60 dalším vynikajícím učitelům s nejvyšší přidanou hodnotou, kteří ve skutečnosti odchod ze školství neplánovali a neodešli by ani bez vyplaceného bonusu. Analýzy ukázaly, že nahrazení neefektivních učitelů je nákladově významně efektivnější než pokusy o udržení učitelů s nejvyšší přidanou hodnotou. V tomto výpočtu je ovšem opomíjen potenciální vedlejší efekt zvýšení platů učitelů s nejvyšší přidanou hodnotou, který by mohl v dlouhodobějším pohledu vést k přilákání více talentovaných jedinců do školství nebo zvýšit úsilí učitelů, kteří již ve školách působí. Skutečný pozitivní efekt tohoto opatření vzdělávací politiky by tedy byl pravděpodobně vyšší.

2.2.3 Závěr

Nakonec Chetty a kol. (2014b) diskutují další otázky spojené s možným zaváděním školských politik využívajících odhady přidané hodnoty učitelů do praxe škol a školských distriktů. Jsou s nimi spojeny dva potenciálně významné problémy, které by se při zavádění takových politik musely uspokojivě vyřešit. Prvním z nich je skutečnost, že zavedení systému, který by systematicky používal odhady přidané hodnoty učitelů k hodnocení učitelů, by mohlo vyvolat nechtěné reakce typu soustředění veškeré výuky na přípravu žáků na standardizované testy nebo podvádění při vlastním testování, které by významně snižovaly hodnotu modelových odhadů přidané hodnoty učitele, které jsou na výsledcích žáků v testech postavené (např. Jacob 2005, Neal a Schanzenbach 2010). Velikost takových účinků by bylo možné odhadnout replikací analýz provedených v této studii ve školských distriktech, ve kterých se již učitelé hodnotí na základě jejich modelově odhadované přidané hodnoty. Pokud by se prokázalo, že chování škol a učitelů ve školských distriktech, které už systém hodnocení učitelů na základě odhadu jejich přidané hodnoty v praxi používají, podstatně sníží schopnost modelů odhadovat přidanou hodnotu, bylo by třeba vyvinout metody a nástroje, které budou méně citlivé na takové behaviorální reakce (Barlevy, Neal 2012). Školské distrikty by také mohly využívat údaje o tom, jak trvalá jsou zlepšení skóre žáků ve standar-

dizovaných testech, což by snižovalo význam potenciálních manipulací s jednotlivým testem a umožnilo vytvoření robustnějšího odhadu přidané hodnoty učitele (Jacob, Levitt 2003).

Druhý potenciální problém je spojen s potřebou porovnávat výsledky výzkumu dlouhodobého vlivu přidané hodnoty učitele na životní a pracovní dráhy jejich žáků s výzkumy, které by zjišťovaly dlouhodobý vliv kvality učitelů měřený jinými způsoby, jako je například hodnocení učitelů vedením škol, nebo hodnocení prováděné na základě pozorování práce učitele ve třídě. Takové výzkumy však nebyly, jak uvádějí autoři studie, která byla publikována v roce 2014, dosud provedeny, nicméně metody vyvinuté autory této studie lze upravit pro vyhodnocení těchto dalších způsobů měření kvality učitele. Takový výzkum by umožnil udělat empiricky podložený odhad optimální váhy dostupných metod hodnocení kvality učitele.

Obecněji řečeno, existuje mnoho aspektů analýzy možných dlouhodobých dopadů práce učitelů, které by ještě měly být prozkoumány, a které by byly užitečné pro navrhování vzdělávací politiky. V této studii byl identifikován vliv jediného učitele na dlouhodobé výsledky jeho bývalých žáků v mladém dospělém věku. Nabízí se otázka, jak by výsledky vypadaly, kdyby se zkoumal vliv více různých učitelů, kteří žáky během jejich školní docházky učili. Sčítá se jednoduše vliv jednotlivých učitelů na dlouhodobé výsledky žáků v jejich mladém dospělém věku, nebo je střídání různě kvalitních učitelů u jednoho žáka ve svém důsledku komplikovanější? Bylo by rovněž užitečné se soustředit nikoliv jenom na „průměrného žáka“, jako v naší studii, ale zkoumat i to, zda jsou někteří učitelé zvláště efektivní (nebo zvláště neefektivní) při práci s žáky, kteří si ve standardizovaných testech vedou špatně, nebo naopak velmi dobře. Nakonec je třeba se zamyslet nad tím, jaké jsou interakce mezi kvalitou učitelů a zájmem rodičů žáků o kvalitu učitelů. Jaké existují motivace rodičů žáků zajistit či udržet učitele s nejvyšší možnou přidanou hodnotou pro své děti a jsou mezi nimi i motivace čistě ekonomické, které by vedly rodiče k rozhodnutí k ochotě přispět takovému učiteli ke zvýšení jeho příjmu? Je každopádně jisté, že důležitou lekcí z této studie je, že zlepšování kvality výuky – ať už s použitím metod hodnocení učitelů prostřednictvím modelových odhadů přidané hodnoty učitelů nebo s použitím jiných metod – pravděpodobně bude přinášet zásadní ekonomické a sociální výhody pro celou společnost.

2.3 Replikace CFR-I a CFR-II měření

Rothstein (2014, 2017a,b) se na datech ze Severní Karolíny pokusil o replikaci empirických zjištění Chettyho a kol. (2014a,b). Využil přitom jimi navržené postupy a CFR modely. I když se mu podařilo úspěšně replikovat všechny analýzy a klíčové výsledky obou uvedených studií, jeho další šetření jej dovedla k závěru, že ani data ze Severní Karolíny, tak ani data z New Yorku nepodporují podstatné závěry obou studií, a to že zkreslení měření přidané hodnoty učitelů jsou zanedbatelná, a že vliv učitelů na jejich žáky je dlouhodobý (Rothstein, 2017a,b). Rothstein (2017a,b) se domnívá, že obavy vzbuzují především tyto tři aspekty a předpoklady CFR metod, které Chetty a kol. (2014a,b) vyvinuli a v praxi použili, a to: (1) způsob, jakým je zacházeno s chybějícími daty, (2) nevhodnost kvazi-experimentální design a (3) nízká robustnost odhadů dlouhodobých efektů učitelů na jejich žáky.

Rothstein (2014, 2017a,b) reaguje na snahu Chettyho a kol. (2014a), kteří se ve své první studii snaží změřit zkreslení míry přidané hodnoty učitele a docházejí k závěru, že zkreslení je minimální. Rothstein (2010, 2014) se však

domnívá, že míra zkreslení nemusí být zanedbatelná, pokud pozorované charakteristiky žáků, které jsou v analýze zahrnuty jako kontrolní proměnné (zejména skóre žáků v minulých testech), plně neabsorbují neměřené determinanty shody mezi žákem a učitelem, které často závisí na požadavcích rodičů nebo na specializaci učitelů.

Zároveň Rothstein (2014, 2017a,b) reaguje i na závěry druhé studie Chettyho a kol. (2014b), ve které se autoři snažili zjistit, zda je přidaná hodnota učitele užitečným ukazatelem jeho vlivu na dlouhodobější výsledky jeho žáků včetně ukončení střední školy, zápisu na vysokou školu a výdělku v mladém dospělém věku a dochází k závěru, že učitelé s vysokou přidanou hodnotou mají dramaticky lepší účinky na všechny tyto výsledky, z čehož vyplývá, že nahrazení učitele s nízkou přidanou hodnotou za jinak podobného učitele s vyšší přidanou hodnotou by přineslo značné výhody pro dlouhodobý úspěch jeho žáků. Rothstein (2014, 2017a,b) se však domnívá, že změnu učitele není možné považovat za validní kvazi-experimentální design. Rothstein dospěl k závěru, že změna průměrné přidané hodnoty všech učitelů daného ročníku v dané škole z roku na rok není nezávislou exogenní proměnnou, která by kvalitou odpovídala náhodně distribuované nezávislé proměnné v experimentu, ale je ve skutečnosti spíše korelovaná s předem danými charakteristikami žáků, které mají prediktivní vliv na výsledky testů.

2.4 Literatura

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (No. w20657). National Bureau of Economic Research. <https://www.nber.org/papers/w20657>

Baker, E. L., P. E. Barton, L. Darling-Hammond, E. Haertel, H. F. Ladd, R. L. Linn, D. Ravitch, R. Rothstein, R. J. Shavelson, and L. A. Shepard. 2010. Problems with the Use of Student Test Scores to Evaluate Teachers. *Economic Policy Institute*. Briefing Paper 278. <https://eric.ed.gov/?id=ED516803>

Barlevy, G., & Neal, D. (2012). Pay for percentile. *American Economic Review*, 102(5), 1805–1831.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Who leaves? Teacher attrition and student achievement* (No. w14022). National Bureau of Economic Research. <https://www.nber.org/papers/w14022>

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, 92(5–6), 1352–1370.

Corcoran, S. P. 2010. *Can Teachers be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice*. Report for the Annenberg Institute for School Reform at Brown University, Education Policy for Action Series. <https://eric.ed.gov/?id=ED522163>

Goldhaber, D., & Chaplin, D. (2012). *Assessing the Rothstein Test: Does It Really Show Teacher Value-Added Models Are Biased?* Working Paper 5. Mathematica Policy Research, Inc. <https://www.tandfonline.com/doi/abs/10.1080/19345747.2014.978059>

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying Effective Teachers Using Performance on the Job. The Hamilton Project Policy Brief No. 2006–01. *Brookings Institution*. <https://www.brookings.edu/articles/identifying-effective-teachers-using-performance-on-the-job/>

Hanushek, E. A. (2009). Teacher Deselection. In Goldhaber D., Hannaway J. (Eds.) *Creating a New Teaching Profession*, 165–80. Washington, DC: Urban Institute Press. https://www.researchgate.net/profile/Eric-Hanushek/publication/265659426_Teacher_Deselection/links/5514277b0cf283ee0834ac05/Teacher-Deselection.pdf

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education review*, 30(3), 466–479.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American economic review*, 100(2), 267–271.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9), 2633–2679.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the impacts of teachers: Reply. *American Economic Review*, 107(6), 1685–1717.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of public Economics*, 89(5–6), 761–796.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3), 843–877.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research. <https://www.nber.org/papers/w14607>

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Research Paper. MET Project. Bill & Melinda Gates Foundation. <https://files.eric.ed.gov/fulltext/ED540959.pdf>

Kane, T. J., Staiger, D. O., & Bacher-Hicks, A. (2014). *Validating Teacher Effect Estimates Using Between School Movers: A Replication and Extension of Chetty et al.* Unpublished.

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, 3(2), 333–362.

- Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and policy*, 6(1), 18–42.
- Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2), 497–532.
- Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2), 263–283.
- Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education finance and policy*, 4(4), 537–571.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1), 175–214.
- Rothstein, J. (2012). *Teacher quality policy when supply matters* (No. w18419). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w18419/w18419.pdf
- Rothstein, J. (2014). Revisiting the impacts of teachers. *UC-Berkeley Working Paper*. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=23310fd91a808eba6fa1d30cdd0857ba18e17af9>
- Rothstein, J. (2017a). Measuring the impacts of teachers: Comment. *American Economic Review*, 107(6), 1656–1684.
- Rothstein, J. (2017b). Supplement to “Revisiting the impacts of teachers. Available July, 24, 2017. https://eml.berkeley.edu/~jrothst/CFR/rothstein_CFR_supplement_jan2017.pdf

3. Experimentální porovnání tří metod hodnocení výkonu učitelů²⁴

Výzkumníci po celá desetiletí dokumentovali vysokou míru heterogenity výsledků žáků (např. Gordon a kol., 2006, Jacob a Lefgren, 2005, Kane a kol., 2008, McCaffrey a kol., 2004, Rivkin a kol., 2005, Rockoff, 2004). Tato zjištění podnítila zástupce decizní sféry k přijímání opatření k odměňování učitelů na základě jejich výkonů. Přitom výkon učitele byl měřen třemi principiálně různými metodami: (1) prostřednictvím odhadů přidané hodnoty učitelů na základě výsledků jeho žáků v testech, (2) z výkonu učitele zaznamenaného v průběhu hospitací (pozorování učitele) ve výuce a (3) z hodnocení učitele jeho žáky za pomoci žákovských dotazníků.

Mezi lety 2008 a 2014 téměř všechny státy USA reformovaly svoji politiku hodnocení učitelů tak, aby hodnocení učitele zahrnovalo alespoň jeden z výše uvedených výkonových ukazatelů (Minnici, 2014). Paralelně se zaváděním těchto opatření byly vznášeny legitimní dotazy na to, zda zaváděné způsoby hodnocení učitelů opravdu odrážejí kauzální efekty jejich pedagogické práce, nebo pouze odrážejí neměřené charakteristiky žáků, které učitelé učí. Přitom za nejkontroverznější byly považovány systémy hodnocení učitelů vycházející z měření výsledků jejich žáků v testech, i když předmětem diskuzí byly i způsoby hodnocení učitelů založené na pozorování učitelů při výuce a metody založené na hodnocení učitelů jejich žáky prostřednictvím dotazníkových šetření. Je zřejmé, že v případě, že některý z těchto způsobů měření explicitně neodráží skutečné výkony učitelů, je místní a školní politika, která na nich staví hodnocení učitelů, chybná.

Studie Bacher-Hickse, China, Kanea a Staigera (2019) na vzorku učitelů matematiky ve čtvrté a páté třídě testuje schopnost tří výše zmíněných běžně používaných způsobů hodnocení učitelů validně predikovat výkon učitele. Během tří školních let byly počítány ukazatele kvality učitele na základě výsledků testů jeho žáků (přidaná hodnota učitele), zároveň ukazatele kvality, které hodnotily učitele na základě pozorování jeho výuky, a ukazatele kvality založené na výsledcích dotazníkových šetření o tom, jak práci učitele vnímají jeho žáci. V prvních dvou letech studie byla shromažďována data všemi třemi způsoby za běžných podmínek. Ve třetím roce studie byli učitelé v rámci svých škol náhodně přiřazeni do tříd a poté byla znovu shromažďována data o výkonu učitelů pomocí všech tří metod. Prostřednictvím shromážděných dat se výzkumníci snažili najít odpovědi na dvě otázky týkající se prediktivní validity výše zmíněných tří způsobů hodnocení práce učitelů:

- Do jaké míry předpovídají skóre, která učitelé získali prostřednictvím třech různých způsobů měření jejich výkonu během prvních dvou let, tedy v době, kdy učitelé nebyli náhodně přiděleni žákům, skórum, které učitelé získali hodnocením jejich výkonu ve třetím roce po jejich náhodném přidělení do tříd?
- Do jaké míry je prediktivní validita každého ze způsobů měření citlivá na základní charakteristiky žáků?

²⁴ Tento text je podrobným výtahem z článku: Bacher-Hicks, A., Chin, M. J., Kane, T. J., Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73, 101919. Z tohoto textu je přebíráán i odkazový aparát.

Pokud neměřené charakteristiky žáků vedou k rozdílu mezi učiteli během prvních dvou let, autoři neočekávají, že výkon z předchozího období bude předpovídat výkon po náhodném přiřazení učitelů do tříd. Kromě toho bude síla tohoto vztahu ilustrovat míru zkreslení odhadů výkonů učitelů měřených v přirozeně se vyskytujících podmínkách (tj. když třídy nejsou náhodně přiřazeny)²⁵.

3.1 Dosavadní studie zaměřené na prediktivní validitu metod hodnocení práce učitelů

Dosud se studie prediktivní validity zaměřovaly převážně na metody odhadující přidanou hodnotu učitele. Ve dvou experimentálních studiích náhodně přidělili učitele do různých tříd v rámci škol, aby posoudili prediktivní validitu různých způsobů měření efektivity učitelů prostřednictvím výsledků standardizovaného testování žáků.

V první studii Kane a Staiger (2008) zjistili, že v každém z případů, který hodnotili, byla při náhodném přiřazení žáků učitelům přidaná hodnota učitele statisticky významným prediktorem skóre testů jejich žáků, ale že pouze modely kontrolující výsledky žáků před tímto náhodným přiřazením generovaly nezkreslené předpovědi. V rámci projektu *Measures of Effective Teaching (MET)* Kane, McCaffrey, Miller a Staiger (2013) realizovali druhý experiment náhodného přiřazení žáků učitelům v rámci školy. Ve své studii použili skóre žáků ve státních testech a testech speciálně vyvinutých v rámci jejich projektu, metodu založenou na pozorování učitele ve třídě a metodu založenou na dotazníkovém šetření mezi žáky, aby vytvořili agregátní ukazatel vlivu učitelů na výsledky testů žáků. Autoři zjistili, že jejich agregátní ukazatel je nezkresleným prediktorem výsledků žáků v testech, které se uskutečnily až po náhodném přiřazení žáků učitelům. Na rozdíl od studie Bacher-Hickse, China, Kanea a Staigera (2019) však Kane a kol. (2013) nebyli schopni vytvořit agregátní ukazatel, který by měřil vliv učitelů na další klíčové výsledky, jako jsou pozorování ve třídě a žákovské dotazníky, kvůli omezením dat autoři nemohli testovat zkreslení odhadu budoucího výkonu učitelů při použití těchto způsobů měření efektivity učitelů²⁶.

Glazerman a Protik (2015) provedli zatím jedinou studii, která pomocí náhodného přidělování učitelů mezi školami testovala prediktivní validitu odhadů přidané hodnoty učitelů v předpovídání výkonu žáků napříč školami. U učitelů základních škol se jejich výsledky shodují s předchozími studiemi v rámci školy, ale při odhadování vlivu učitele středních škol nacházejí důkazy o značném zkreslení.

Kromě výsledků studií používajících randomizované experimenty se tři nedávné studie spoléhaly na kvazi-ex-

25 Jednou z obav je, že „přirozeně se vyskytující prostředí“ vzorku tříd čtvrtého a pátého ročníku při výuce matematiky nemusí dobře reprezentovat jiné ročníky a předměty. Například Steinberg a Garrett (2016) ukazují, že skóre změřená pozorováním učitelů matematiky méně závisí na výchozích výsledcích jejich žáků, než je tomu u učitelů angličtiny a dalších předmětů. Pokud by byli žáci již v letech před experimentem náhodně přiřazováni k učitelům, test by nebylo možno zobecňovat na prostředí, ve kterých jsou žáci systematicky přiřazováni k učitelům. V dalších částech textu autoři dokumentují rozsah předrandomizačního třídění v jejich vzorku a ukazují, že je v souladu s Kanem a kol. (2013), kteří se opírají o vzorek tříd ve čtvrtém až osmém ročníku. Nicméně diskutují o omezeních svého vzorku a v závěru poskytují směr pro budoucí výzkum.

26 Kane a kol. (2013) shromáždili údaje o výsledcích žáků v testech za tři roky, ale údaje o pozorováních ve třídě a dotaznících mezi žáky získali pouze za dva roky. Vzhledem k tomu, že jejich odhad kvality učitelů vyžaduje tři roky dat, konstruují svůj agregátní ukazatel tak, aby byl odhadem budoucího dopadu učitelů na výsledky testů žáků, nikoli odhadem budoucího vlivu jiných ukazatelů, jako jsou pozorování ve třídě a dotazníkové šetření mezi žáky. Vzhledem k tomu, že tato studie trvá tři školní roky, autoři této studie získali různé odhady pro předpovídání vlivu učitelů ve třech dimenzích měřených: prostřednictvím výsledků žáků v testech, pozorováních ve třídě a prostřednictvím odpovědí v dotazníkových šetřeních se žáky.

perimentální design, aby otestovaly prediktivní validitu odhadů přidané hodnoty učitele. Tato kvazi-experimentální metoda – představená Chetty, Friedmanem a Rockoffem (2014) – předpovídá změny ve výsledcích žáků v testech pomocí přirozeně se vyskytujících exogenních variací v přiřazování učitelů do tříd, jak učitelé přecházejí ze školy do školy a z ročníku do ročníku. Použitím této metody Chetty a kol. (2014) zjistili, že skóre přidané hodnoty učitelů, při kontrole výsledků žáků před vstupem do tříd vyučovaných těmito učiteli, byly nezkreslenými prediktory výsledků žáků. Bacher-Hicks, Kane a Staiger (2014) replikovali tuto kvazi-experimentální metodu v Los Angeles s podobnými výsledky. Také Rothstein (2017) našel málo důkazů o zkreslení predikcí při aplikaci stejné techniky na datech ze Severní Karolíny, ale tvrdil, že výsledky nemusí dodat dostatečnou informaci, protože kvazi-experiment selhal v placebo testu²⁷.

Na rozdíl od pozornosti, které byla v odborné literatuře věnovaná měření přidané hodnoty učitele, byla relativně málo zkoumána prediktivní validita jiných způsobů měření kvality učitelů používaných vedoucími školských distriktů a řediteli škol při hodnocení učitelů. Blazar (2018) hodnotil prediktivní validitu vlivu učitele na chování žáků, jejich efektivitu a štěstí (tato data byla získána prostřednictvím sebe-evaluačních dotazníků). Měření vlivu učitele na tyto charakteristiky žáků obsahuje užitečné prediktivní informace o jeho budoucím vlivu po náhodném přiřazení učitele žákům, výsledky těchto predikcí jsou ovšem méně přesné než studie využívající přidanou hodnotu založenou na testech, takže je v těchto případech obtížné kvantifikovat velikost zkreslení.

Garrett a Steinberg (2015) na základě dat z projektu MET zjistili, že výsledky pozorování učitelů ve třídě jsou pozitivními prediktory testových výsledků žáků po náhodném přidělení žáků učitelům, ale kvůli charakteru dat existuje jen omezená možnost hodnotit jejich prediktivní validitu. Steinberg a Garrett (2016) podobně spoléhají na data MET při měření vztahu mezi výsledky žáků před přidělením do tříd a následným skóre efektivnosti práce jejich učitelů založeném na pozorování ve třídě. Potvrzují dřívější neexperimentální zjištění (např. Whitehurst, Chingos a Lindquist, 2014) tím, že ukazují, že učitelé, kteří jsou náhodně přiděleni do tříd s žáky, kteří mají lepší průměrné výsledky testů před experimentem, získávají vyšší skóre v měřeních efektivnosti založených na pozorování ve třídě, což platí zejména pro učitele angličtiny a učitele odborných předmětů. I když přímo neodhadují velikost zkreslení odhadu, tyto výsledky naznačují, že měření prostřednictvím pozorování ve třídě mohou být ovlivněna charakteristikami žáků.

Stručně řečeno, řada nedávných studií hodnotila potenciální zkreslení při používání různých metod hodnocení učitelů, ale primárně prokázaly nezkreslenost pouze u jednoho typu metod: měření vlivu učitelů prostřednictvím analýzy výsledků testů jejich žáků. K dnešnímu dni pouze jedna studie zkoumala zkreslení předpovědí v odhadech výkonu učitelů na základě měření založených na dotazníkových šetřeních mezi žáky, zároveň neexistují žádné přímé důkazy o nezkreslenosti měření založených na hodnocení učitelů prostřednictvím pozorování ve třídách – metrika, kterou vedoucí školských distriktů a ředitelé škol nejčastěji používají k hodnocení učitelů.

²⁷ Rothstein (2017) nachází jen málo důkazů o zkreslení při použití metody přidané hodnoty učitele, když replikoval studii dle specifik uvedených v Chetty a kol. (2014). Rothstein však tvrdí, že kvazi-experiment sám o sobě není platným testem, protože selže v testu placebo, spočívající v korelaci změny v přidané hodnotě se změnami ve skóre předchozích testů. Chetty, Friedman a Rockoff (2017) na tuto kritiku reagují argumentem, že placebo test je ovlivněn spíše mechanickým efektem než selekčním zkreslením.

3.2 Design výzkumu

K vývoji a testování prediktivní validity různých způsobů měření efektivity učitelů Bacher-Hicks, Chin, Kane a Staiger (2019) použili data shromážděná v rámci studie *National Center for Teacher Effectiveness* (NCTE). Studie zahrnovala čtyři velké školní distrikty ve státech na východním pobřeží USA a trvala tři školní roky, od roku 2010, resp. 2011 až do školního roku 2012/2013. Během všech tří školních roků byly shromažďovány údaje o učitelích a žácích. Ve třetím roce byli způsobilí a zúčastnění učitelé navíc náhodně zařazeni do tříd. Ve všech čtyřech školních distriktech 316 učitelů čtvrtého a pátého ročníku souhlasilo s účastí v alespoň jednom ze tří ročníků studie a 66 učitelů bylo způsobilých a souhlasilo s účastí na náhodném přidělení učitelů do tříd ve třetím ročníku studie²⁸.

NCTE shromáždila již existující administrativní data, včetně informací o jednotlivých třídách, demografických informací a výsledcích státních testů pro všechny žáky čtvrté a páté třídy ve čtyřech participujících školních distriktech. Kromě toho byly v rámci studie shromažďovány informace o žácích a jejich učitelích, kteří pracovali ve třídách participujících na šetření, a to: výkon žáka v testu matematiky, který byl vyvinut v rámci tohoto projektu (výsledek neměl žádný vliv na hodnocení žáků); odpovědi žáků v dotazníkovém šetření; digitálně nahrané hodiny matematiky, které sloužily jako podklad pro hodnocení výuky ve třídě; a odpovědi učitelů na dotazník zjišťující informace o přípravě na vyučování a další charakteristiky učitelů.

3.2.1 Měření

Autoři použili již existující administrativní záznamy a další data shromážděná NCTE ke generování odhadů výkonu učitelů na základě pěti měření:

- skóre žáků ve státních standardizovaných testech z matematiky;
- výsledky žáků v matematickém testu vytvořeném v rámci projektu (Hickman, Fu a Hill, 2012);
- výkon učitelů měřený nástrojem *Mathematical Quality of Instruction*, který byl vyvinutý pro hospitace ve výuce (Hill a kol., 2008);
- výkon učitelů měřený nástrojem *Classroom Assessment Scoring System* (Pianta, La Paro, Hamre, 2008);
- informace z žakovského dotazníku zaměřeného na vnímání dění ve třídě – *Tripod-based perception survey* (Ferguson, 2009).

²⁸ K menšímu počtu učitelů účastnících se experimentu ve třetím roce této studie přispělo několik faktorů. Nejdůležitějším byla fluktuace učitelů v průběhu času. Z 316 učitelů, kteří se zúčastnili některého ze tří ročníků studie, zůstalo ve třetím roce studie vyučovat na zúčastněných školách pouze 132. Tak vysoká míra fluktuace učitelů není atypická, zejména v městských distriktech (Papay, Bacher-Hicks, Page a Marinell, 2017). Dále se v některých školách změnilo vedení, což způsobilo, že jen menší počet ředitelů měl nadále zájem účastnit se experimentu, v jehož rámci mělo dojít k náhodnému přidělování žáků a učitelů. Ze 132 učitelů nakonec v projektu zůstalo 78 učitelů. Ti měli zájem o účast v experimentu a zároveň byli ve školách, ve kterých vedení projevilo zájem se experimentu zúčastnit. Z těchto 78 učitelů 66 učitelů splnilo všechny ostatní podmínky způsobilosti (např. museli učit třídu s nejméně pěti žáky, u kterých byly k dispozici aktuální a výchozí výsledky standardizovaných testů, nesměli být jediným učitelem v bloku náhodných přidělení atd.).

Výsledky žáků ve státních standardizovaných testech pocházely ze tří různých testů. Mezi testy byla značná variabilita formátu a kognitivní náročnosti položek. Například žáci v jednom okrese dělali test, který byl kompletně složen z položek s vícenásobnou volbou (*multiple choice items*), zatímco žáci v jiném distriktu dělali test s otevřenými položkami, které byly výrazně obtížnější (Lynch, Chin a Blazar, 2017). Aby autoři studie zohlednili tyto rozdíly mezi testy, přeškálovali výsledky žákovských testů dle okresů, ročníků a akademických roků pomocí van der Waerdenových standardizačních metod založených na pořadí úspěšnosti v testech (Conover, 1999).

Ve spolupráci s *Educational Testing Service* byl NCTE vyvinut test z matematiky pro čtvrtou a pátou třídu, který byl navržený tak, aby byl v souladu se základními státními standardy pro matematiku (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) a s dalšími projektovými nástroji pro hodnocení kvality učitelů. Testy obsahovaly kromě tradičních položek s vícenásobnou volbou také položky se strojově čitelnými možnostmi numerických odpovědí na otázky a otevřené položky. Podobně jako výsledky státních testů autoři standardizovali výsledky speciálně vytvořených testů tak, aby měly průměr nula a směrodatnou odchylku jedna v rámci ročníku a školního roku. Protože testy byly napříč distrikty stejné (na rozdíl od skóre státního testu), autoři zachovali rozdíly v průměrech a rozptylech mezi jednotlivými distrikty.

Měření pozorování učitelů pomocí nástrojů MQI a CLASS byla založena na až třech video-záznamech vyučovacích hodin matematiky na jednoho učitele za rok. Učitelé si sami vybírali, které hodiny zaznamenají, a to pod podmínkou, že si k záznamu vyberou hodiny, které jsou typické pro jejich specializaci²⁹. Nahrávané hodiny byly v průměru šedesátiminutové. Výukové hodiny zachyceny na videích byly bodovány (hodnoceny) v souladu se dvěma zavedenými pozorovacími nástroji: MQI (Hill a kol., 2008) a CLASS (La Paro a kol., 2012). Nástroj MQI byl navržen tak, aby posuzoval odbornost učitelů a jejich schopnost poskytovat kvalitní výuku matematiky. Tento nástroj se skládá ze 14 kódovaných částí, jejich prostřednictvím je možno hodnotit výukovou praxi, viz například: zda se používá více přístupů k řešení problému; zda se používá správný matematický jazyk a zápis; zda dochází k nápravě obtíží žáků v matematice (Hill a kol., 2008). Naproti tomu nástroj CLASS měří všeobecnou pedagogickou způsobilost učitele nezávisle na vyučovaném předmětu. Tento nástroj má 12 kódovaných částí a jeho prostřednictvím lze hodnotit charakteristiky výuky učitele ve třídě jako je například schopnost učitele dosahovat ve třídě pozitivní klima; schopnost učitele poskytovat kvalitní zpětnou vazbu; dovednosti učitele usměrňovat chování žáků; nebo schopnost učitele zapojovat žáky do výuky (La Paro a kol., 2012).

Aby Bacher-Hicks, Chin, Kane a Staiger (2019) vygenerovali roční celkové skóre založené na pozorování metodou MQI u jednotlivých učitelů, nejprve pro každého učitele zprůměrovali skóre získané v jednotlivých částech v rámci jednotlivých hodin a poté udělali průměr z měření získaných v jednotlivých hodinách během jednoho roku. Protože nástroj byl stejný ve čtvrtém a pátém ročníku a napříč školskými distrikty, standardizovali tato skóre na úrovni učitelů tak, aby měla průměr nula a směrodatnou odchylku jedna v rámci školního roku. Stejným postupem vypočítali skóre pozorování učitele metodou CLASS³⁰.

29 Ho a Kane (2013) zjistili, že hodnocení učitelů na základě pozorování ve třídě bylo stabilní mezi vyučovacími hodinami, které si učitelé sami vybrali, a ostatními hodinami.

30 Autoři standardizovali měření realizovaná pomocí metod MQI a CLASS na úrovni učitelů, protože data na úrovni žáků neexistují; standardizovali míry odvozené od výsledků státních testů žáků, výsledků testů vytvořených v projektu a odpovědí v dotazníkovém šetření na úrovni žáků.

Nakonec odvodili ukazatel z odpovědí žáků v dotazníkovém šetření, který zahrnoval 26 položek Likertovy škály (Ferguson, 2009). Položky zjišťovaly názory žáků na řadu témat, například zda se žáci domnívají, že je matematika jejich učitelem prezentovaná poutavě; zda mají pocit, že se učitel o ně stará; jestli učitel vyzývá žáky, aby se zapojili do výuky matematiky; jestli učitel jasně prezentuje matematický obsah; jestli učitel pravidelně hodnotí porozumění probírané látce; jestli učitel žákům poskytuje užitečnou zpětnou vazbu; nebo zda je třída produktivní během celé vyučovací hodiny matematiky. Následně autoři vypočítali průměrné skóre z odpovědí ze všech 26 položek a standardizovali tato skóre na úrovni žáků tak, aby měla průměr nula a směrodatnou odchylku jedna během jednoho školního roku.

Pro tři měření, která vycházela z dat získaných od žáků (tj. skóre státních testů, skóre projektových testů a odpovědi dotazníkovém šetření) autoři vygenerovali odhad kvality učitele jako průměrného rezidua na úrovni učitele v jednotlivých letech z následující regresní rovnice OLS:

$$A_{i,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{k,t}\delta + \varphi_{g,t} + \eta_d + \varepsilon_{i,k,t}, \quad (1)$$

Kde $A_{i,k,t}$ je standardizovaný výsledek státního testu pro žáka i , kterého učil učitel k během školního roku t . Kromě fixních efektů po jednotlivých ročnících, $\varphi_{g,t}$ a fixních efektů distriktu η_d autoři zahrnuli následující kontrolní proměnné: $A_{i,t-1}\alpha$ kubický polynom výsledků testu žáka i v roce předcházejícím roku t ; $S_{i,t}\beta$ vektor ukazatelů pro pohlaví, rasu a etnický původ, způsobilost k bezplatnému obědu nebo obědu za sníženou cenu (způsobilost FRPL), omezenou znalost angličtiny a status speciálního vzdělávání; a $P_{k,t}\delta$ vektor průměrných charakteristik vrstevníků žáka i ve stejné třídě, včetně průměrného skóre testu předcházejícího experimentu a průměrů $S_{i,t}$ na úrovni třídy.³¹ Chyba měření na úrovni žáka je $\varepsilon_{i,k,t}$. Jak je typické v modelech s přidanou hodnotou, autoři odhadli roční rezidua učitele zprůměrováním $\varepsilon_{i,k,t}$ napříč žáky učitele k v roce t , což poskytuje odhad výkonu učitele k v roce t . Analogickým postupem generovali rezidua pro každého učitele pro každý rok měření jeho výkonu změřeného z výsledků testu, který byl pro jejich žáky speciálně vytvořený v rámci projektu. Stejný postup byl použit i pro rezidua získaná z měření prostřednictvím výsledků dotazníkového šetření mezi žáky. V obou případech to autoři dělali za využití rovnice (1) změnou závislé proměnné na standardizované skóre pro test vytvořený v rámci projektu nebo standardizované skóre vypočítané z dat z dotazníkového šetření mezi žáky.

Protože pro tato dvě měření neexistovala data na úrovni žáků, vygenerovali pro tato měření rezidua za každý rok učitele pomocí následující regresní rovnice OLS:

$$M_{k,t} = P_{k,t}\delta + \eta_d + \varepsilon_{k,t}, \quad (2)$$

kde $M_{k,t}$ je míra učitelova skóre vypočítaná na základě pozorování ve třídě a $P_{k,t}\delta$ je stejný vektor průměrných charakteristik vrstevníků žáka i použitý v rovnici (1), která zahrnovala průměrné výsledky testu předcházejícího výzkumu a průměry charakteristik žáka, $S_{i,t}$. Tento model autoři odhadli samostatně pro standardizované skóre MQI a CLASS ve třídě, aby vytvořili dvě rezidua pro každého učitele za každý rok. Tato rezidua byla použita k predikci výkonu učitele v roce náhodného přiřazení žáků učitelům.

31 Rovnice (1) představuje úplný soubor kontrolních proměnných, ale autoři odhadují další specifikace s menším počtem kontrolních proměnných, aby odpověděli na druhou výzkumnou otázku (viz specifikace v tabulkách 6 a 7).

3.2.2 Popis experimentu s náhodným přiřazením učitelů a žáků

Během druhého roku sběru dat (2011/2012) pracoval projektový tým NCTE se zaměstnanci participujících škol na identifikaci učitelů, kteří splnili nezbytné požadavky, aby byli součástí projektu ve třetím roce (2012/2013). Přitom byly stanoveny základní podmínky způsobilosti pro účast učitelů v další části projektu, ve které mělo dojít k náhodnému přiřazení žáků učitelům, a to: (a) učitelé museli být součástí skupiny dvou nebo více učitelů projektu NCTE, kteří na dané škole učili ve stejném ročníku, a (b) ředitelé museli na učitele v této skupině pohlížet jako na schopné bez větších úprav vyučovat podle jakéhokoliv ze školních plánů³².

Učitelé splňující tyto dvě podmínky byli zařazeni do randomizačního bloku po dvou nebo třech učitelích. Ředitelé škol vygenerovali seznam žáků ve třídách učitelů v každém bloku randomizace. Pokud by například randomizační blok měl tři učitele, ředitelé školy by sestavili tři seznamy žáků. Poté, co ředitelé škol vytvořili tyto seznamy, byly předloženy studijnímu týmu NCTE, který náhodně přiřadil způsobilé učitele k třídám a poté vrátil odpovídající seznamy ředitelům škol. Z celkového počtu 29 randomizačních bloků 21 obsahovalo dva učitele a 8 obsahovalo tři učitele.

Tabulka 1: Shrnutí informací o náhodně přiřazených žácích

	Number of students	Percent of total
Remained with randomly assigned teacher	838	71.2
Switched teacher within random assignment block	50	4.2
Switched teacher within school	148	12.6
Left school or district	141	12.0
Number of students	1177	100

Poznámka: Vzorek tvoří žáci čtvrté a páté třídy ze čtyř školských distriktů, kteří byli náhodně přiděleni do třídy ve školním roce 2012/13.

V ideálním případě by každého náhodně přiděleného žáka učil učitel, kterému byli náhodně přiděleni. Avšak vzhledem k tomu, že tyto seznamy byly sestaveny před začátkem roku s náhodným přidělením, určitý pohyb byl nevyhnutelný. V tabulce 1 jsou dokumentovány informace o 1177 žácích v náhodném výběru. Je pozoruhodné, že 71 procent těchto žáků zůstalo ve své náhodně přidělené třídě po celý školní rok, což zlepšuje shodu ve srovnání s předchozími randomizačními studiemi. Například v šesti lokalitách popsanych ve studii Kane a kol. (2013), zůstalo ve své náhodně přidělené třídě mezi 27 procent a 66 procent žáků.

³² V některých případech bylo požadováno, aby někteří žáci byli spárováni s konkrétními učiteli v rámci randomizačního bloku (např. pokud byl pouze jeden učitel certifikován pro výuku žáků s omezenou znalostí angličtiny). V těchto případech pracovníci NCTE umožnili těmto žákům, aby byli nenáhodně spárováni s příslušným učitelem, a poté náhodně obsadili zbývající místa ve třídě. Tyto nenáhodně umístěné žáky autoři vyloučili ze všech analýz, které jsou omezeny na vzorek vzniklý náhodným přiřazením, ale zahrnují je při generování agregovaných proměnných charakterizujících „peer control“. Přibližně 7 procent žáků ve třídách s náhodným výběrem bylo spárováno nenáhodně.

3.3 Empirická strategie

Empirická strategie zahrnuje dva klíčové kroky, které jsou analogické se stávajícími metodami používanými v předchozích studiích kvality učitelů (např. Chetty a kol., 2017, Kane a Staiger, 2008, Kane a kol., 2013). Nejprve jsou vygenerovány odhady očekávaného výkonu učitelů na základě dat pocházejících ze dvou let, které předcházely samotnému experimentu s náhodným přiřazováním žáků a učitelů. Za druhé je posuzováno, zda tyto neexperimentální odhady přesně předpovídají výkon učitele po náhodném přiřazení.

3.3.1 *Predikce očekávání výkonu učitele*

Cílem tohoto kroku je předpovědět výkon učitelů v roce náhodného přiřazení pouze pomocí neexperimentálních informací z předchozích let. K vytvoření této predikce se používají rezidua spočítaná pro každého učitele a každý ze dvou roků předcházejících roku náhodného přiřazení (jak je popsáno v rovnicích (1) a (2)). Tím, že se používají data z několika předchozích let, jsou v předpovědích zohledňována meziroční kolísání výkonu učitelů způsobená chybou měření.

3.3.2 *Porovnání očekávaného výkonu se skutečnými výsledky po náhodném přiřazení*

K odhadu vztahu mezi predikcemi výkonu učitelů a skutečnými výsledky po náhodném přiřazení je používán odhad instrumentálních proměnných. Zatímco efektivita učitelů, kterým byli žáci přidělováni náhodně, není v korelaci s pozorovatelnými nebo nepozorovatelnými charakteristikami žáka, nemůžeme si být jisti, že efektivita skutečného učitele, u kterého není přidělování zcela náhodné, jejich charakteristikami skutečně není ovlivněna. Například, jak autoři zdokumentovali v tabulce I, ačkoli procento žáků, kteří zůstali ve třídě s náhodně přiřazeným učitelem, bylo vyšší než v minulých studiích pracujících s náhodným přiřazením, někteří žáci přešli do různých tříd ve škole, což nemuselo být náhodné (např. pohyb mohl následovat po stejném rozřazování žáků do tříd, ke kterému dochází v běžném roce). Efektivita skutečného učitele žáka se měří v porovnání s efektivitou náhodně přiděleného učitele.

V první fázi je odhadován vliv náhodně přiděleného učitele na skutečnou efektivitu učitele. Tento postup je prováděn pro tři různé způsoby měření výkonu učitele na základě dat žáků. Bylo zjištěno, že zvýšení o jednu jednotku standardní odchylky v odhadovaném výkonu učitele bylo spojeno se zvýšením skutečného výkonu učitele o 0,82 až 0,91 jednotky standardní odchylky, a to u všech třech typů měření. Ve druhé fázi jsou predikovány skutečné výsledky žáků ve třetím roce studie, kdy byli žáci a učitelé přiřazováni náhodně.

3.4 Shrnutí hlavních empirických zjištění

Autoři této studie přinášejí do dosavadního výzkumu tři klíčové příspěvky. Za prvé, přispívají k hromadění důkazů o prediktivní platnosti měř přidané hodnoty učitelů. V souladu s předchozími studiemi, které využívají randomizované experimenty, zjistili, že měření přidané hodnoty založené na testech jsou nezkreslenými prediktory vlivu učitelů na výsledky žáků i po náhodném přiřazení učitelů k žákům. I když se tyto experimentální výsledky opírají o méně předpokladů, než bývá obvyklé u kvazi experimentálních testů zkreslení, což je považováno za jejich výhodu, jejich nedostatkem je, že jsou méně přesné. Aby autoři vytvořili přesnější experimentální odhad výkonu učitelů, používají meta-analytické metody ke kombinaci výsledků této studie s výsledky dvou dalších existujících experimentů s náhodným přiřazením žáků učitelům v rámci školy. Experimentální odhad využívající spojená data z více experimentů ukazuje, že přidaná hodnota je validním prediktorem průměrného skóre žáků v testech po náhodném přiřazení žáků učitelů a zároveň je přesnější než existující experimentální důkazy.

Za druhé, autoři studie předkládají nové důkazy o prediktivní validitě metod založených na pozorování ve třídě a dotazníkových šetření mezi žáky. Vzhledem k tomu, že žáci obvykle nejsou k učitelům přidělováni náhodně, efektivita těchto metod může být ovlivněna stejnými výběrovými zkresleními jako častěji kritizovaná měření založená na testech. Autoři této studie zjistili, že skóre výkonnosti učitelů založená na metodě pozorování ve třídách, kde byli učitelé a žáci přiřazeni přirozeným způsobem, předpovídá skóre učitelů změřená touto metodou po jejich náhodném přiřazení do tříd. Konkrétně řečeno, nelze zamítnout hypotézu koeficientu jedna (tj. hypotéza předpokládající, že skóre získaná pozorováním učitelů při přirozeném přiřazení dokonale předpovídá skóre získaná pozorováním po náhodném přiřazení) ani nelze zamítnout nulovou hypotézu (tj. hypotéza předpokládající, že mezi skórem efektivity učitele změřeným pozorováním v přirozeném přiřazení a po náhodném přiřazení žáků učitelům neexistuje žádný vztah, tedy že skóre získaných pozorováním při přirozeném přiřazení žáků a učitelů nemají žádnou prediktivní sílu). Protože jsou však tyto odhady méně přesné než odhady založené na výpočtu přidané hodnoty učitele, autoři udávají, že mohou zamítnout také přiměřenou míru zkreslení odhadu (95 procent interval spolehlivosti se pohybuje od 0,51 do 1,65). U metody měření výkonu učitelů, která byla založena na datech z žakovských dotazníků, autoři studie zamítli hypotézu jedna, ale nedokázali zamítnout nulovou hypotézu, což naznačuje, že odpovědi žáků z dotazníkového šetření obsahují málo informací o výkonech učitelů; tyto odhady výkonu učitelů jsou opět podstatně méně přesné než odhady založené na měření přidané hodnoty učitele.

Za třetí, autoři studie vyhodnocovali citlivost těchto tří měření na vliv kontrolních proměnných. V praxi modely pracující s měřením přidané hodnoty učitele obvykle kontrolují přinejmenším proměnné zachycující výsledky testů žáků, se kterými do třídy vedené učitelem, jehož přidaná hodnota je měřena, nastupovali, zatímco metody založené na pozorování ve třídě obvykle žádné takové kontroly nezahrnují. V souladu s předchozími výzkumy o přidané hodnotě autoři zjistili, že je nezbytné kontrolovat výsledky žakovských testů před započítáním experimentu, aby bylo možné vytvořit nezkreslený prediktor vlivu učitele. Na rozdíl od metody přidané hodnoty učitele však zjistili, že prediktivní síla měření založených na pozorování ve třídách není ovlivněna rozhodnutím, zda v modelech použít či nepoužít kontrolní proměnné změřené na úrovni žáků nebo třídy. I když je tato studie omezena jen na analýzu efektivity učitelů při výuce matematiky ve třídách čtvrtého a pátého ročníku, toto zjištění obecně poskytuje podporu pro měření efektivity učitelů založené na pozoro-

vání ve třídách i v jiných předmětech, protože jsou typicky operacionalizována v podmínkách odpovídajících běžnému způsobu, jakým se v praxi takové hodnocení práce učitelů provádí (tj. stavu, kdy hodnotitelé nepoužívají korekce pro složení školních tříd).

3.5 Důsledky pro formování politik a závěr

Nedávné vzdělávací reformy zaměřené na hodnocení učitelů nevyhnutelně vyvolaly otázky o prediktivní validitě používaných metod měření výkonu učitelů. Zjištění Bacher-Hickse a kol. (2019) jsou nejnovější ze série studií, které naznačují, že nejkontroverznější způsob měření výkonu – měření přidané hodnoty učitele založené na výsledcích testů jejich žáků – předpovídá vliv učitele na výsledky žáků bez zkuslení poté, co jsou žáci a učitelé k sobě náhodně přiřazeni.

Doposud se veřejná polemika (a výzkum prediktivní validity) zaměřovala na odhady přidané hodnoty učitelů založené na výsledcích jejich žáků v testech. Více než dvě třetiny učitelů jsou však v ročnících nebo předmětech, ve kterých se žádné testování neuskutečňuje., kde taková měření přidané hodnoty učitele nelze aplikovat. Výsledkem je, že většina učitelů je hodnocena na základě jiných metod hodnocení, než jsou měření přidané hodnoty učitelů založená na testech – a u těchto dalších způsobů měření se dosud neprokázalo, že by zachycovaly kauzální vliv efektivity učitelů na výsledky žáků. Experiment s náhodným přiřazením učitelů a žáků byly použity ke zkoumání, zda metoda využívající pozorování práce učitele ve třídě a její následné hodnocení – metoda, kterou nejčastěji používají vedoucí pracovníci školských distriktů a ředitelů škol k hodnocení učitelů – a metoda využívající výsledky průzkumů mezi žáky – metoda aplikovaná mnoha vzdělávacími agenturami ve snaze naplnit požadavky federální směrnice – zachycují kauzální dopady práce učitele. I když jsou tyto způsoby měření méně přesné než měření přidané hodnoty učitelů na základě výsledků testů jejich žáků, bylo zjištěno, že pozorování práce učitelů ve třídě shromážděné před experimentem, v době, kdy byli žáci přiřazeni do tříd a k učitelům přirozeným způsobem, obsahují významné prediktivní informace o výkonu učitelů poté, co jsou učitelé a žáci k sobě přiřazeni náhodným způsobem. Na druhou stranu, metody založené na analýze odpovědí žáků v dotazníkovém šetření nepředpovídají výkon učitelů po náhodném přiřazení učitelů a žáků, což naznačuje omezenou prediktivní validitu tohoto způsobu hodnocení práce učitelů.

Byla zkoumána také citlivost těchto výsledků na působení kontrolních proměnných. Částečně kvůli výzkumům provedeným v minulosti je ve většině systémů, které jsou na úrovni jednotlivých států USA používány k měření přidané hodnoty učitele s využitím výsledků testů žáků, během měření používány jako kontrolní proměnné výsledky žáků ve starších testech. Podobně jako jiné studie i studie Bacher-Hickse a kol. (2019) poskytla důkazy, že kontrola skóre v testech předcházejících měřenému období je nezbytná pro vytvoření nezkráceného odhadu přidané hodnoty učitele. Bez podobných důkazů o prediktivní validitě metod hodnocení založených na dotazníkových šetřeních mezi žáky nebo metod využívajících pozorování práce učitelů ve třídě však většina států nemá návod, zda a případně jak by výsledky získané těmito metodami měly být upraveny, aby byl kontrolován vliv průměrných charakteristik a výsledků žáků v každé třídě. Vzhledem k tomu, že téměř žádná instituce provádějící hodnocení učitelů nekoriguje výsledky získané pozorováním ve třídě, důkazy o legitimitě této praxe

jsou zásadní – zvláště u některých studií, které zjistily, že výkon učitelů hodnocený běžně používanými metodami pozorování koreluje se testovými výsledky žáků změřenými před hodnoceným obdobím (např. Steinberg a Garrett, 2016, Whitehurst a kol., 2014). I když se studie Bacher-Hickse a kol. (2019) zaměřuje pouze na výuku matematiky ve vybraných ročnících základních škol, poskytuje první důkazy, které naznačují, že tato běžná praxe může být vhodná: ve zkoumaném výzkumném vzorku nebylo nutné upravovat výsledky o vliv rozdílů v charakteristikách žáků ani výsledcích testů žáků před sledovaným obdobím, aby bylo možno vytvořit nezaujaté odhady měření založené na pozorování práce učitelů ve třídách.

Jedno z hlavních omezení zmiňované studie spočívá v nepřesnosti odhadů validity, zejména u metody založené na odpovědích žáků v dotazníkovém výzkumu. Jak bylo uvedeno výše, bylo to způsobeno – alespoň částečně – neschopností použít jako kontrolní proměnnou odpovědi žáků v dotazníkovém šetření, které by se uskutečnilo ještě před realizací studie (takové šetření se neuskutečnilo). Protože žáci nebyli sledováni longitudinálně, nebylo možno kontrolovat žádné základní proměnné na úrovni žáků, které nebyly obsaženy v administrativních záznamech, které jsou pořizovány standardně a jsou proto k dispozici i za období před uskutečněním studie. Na základě toho mohou budoucí experimenty zvážit dlouhodobé sledování žáků, aby bylo možné kontrolovat vliv proměnných změřených před realizací vlastní studie na hodnoty proměnných zjištěných během realizace studie. Druhým důvodem nepřesnosti je relativně malý vzorek učitelů. Vzhledem k tomu, že přesnost se bude zlepšovat s rostoucí velikostí vzorku, budoucí experimenty mohou zvážit spoléhání se na větší vzorky při testování zkreslení výsledků zjištěných metodou pozorování ve třídě a analýzou odpovědí žáků v dotazníkových šetřeních. Budoucí studie mohou také zvážit použití kvazi-experimentální identifikační strategie Chetty a kol. (2017) při využití velkých administrativních databází výsledků pozorování práce učitele ve třídě nebo odpovědí žáků v šetřeních. Kromě přínosu velkého vzorku se tato kvazi-experimentální strategie opírá o pohyb učitelů napříč školami, což by poskytlo důkaz o tom, do jaké míry je výkon učitelů v těchto měřících specifický pro školu nebo do jaké míry je specifická v určitém kontextu.

Dalším omezením vzorku použitým ve studii Bacher-Hickse a kol. (2019) je, že zahrnoval pouze výuku matematiky ve třídách v základních školách. Výsledky této studie proto nelze zobecňovat na jiné ročníky a předměty. Například Steinberg a Garrett (2016) zjišťují, že skóre zjištěné metodou pozorování učitelů matematiky ve třídách méně souvisí se výsledky žáků ve třídě než pozorování učitelů angličtiny a učitelů v jiných předmětech. To naznačuje, že zkoumaný vzorek může být ze své podstaty méně náchylný ke zkreslení způsobené nezahrnutými proměnnými (*omitted variable bias*) než vzorek z jiných ročníků anebo předmětů. Bohužel nelze vyhodnotit rozsah zkreslení predikcí u efektivnosti učitelů angličtiny a dalších předmětů, ale lze doufat, že budoucí studie – možná opírající se o velké administrativní databáze – bude schopna hodnotit prediktivní validitu u měření získaných pozorováním i v jiných předmětech a ročnících. Budoucí studie budou moci vyhodnotit prediktivní validitu ve specifických pozorovacích nástrojích používaných mnoha školskými distrikty, jako je např. *Danielson Framework for Teaching (FFT)*. Ačkoli dvě metody založené na pozorování použité ve studii Bacher-Hickse a kol. (2019) – CLASS a MQI – se používají méně často, jejich výsledky jsou podobné, což naznačuje, že uvedená zjištění mohou platit pro jiné metody založené na pozorování. Předchozí práce navíc ukázaly, že výkon učitelů měřený různými metodami založenými na pozorování je vysoce korelovaný (Kane, Staiger, 2012), přičemž zejména CLASS vykazuje vysokou empirickou a kvalitativní podobnost s FFT.

Nakonec lze poznamenat, že vlastnosti jakékoli metody měření výkonu učitele by se mohly změnit, protože se tyto metody používají pro stále širší účely a je o ně větší zájem. Ve zmiňované studii nebyl výkon učitelů měřený pozorováním ve třídě a šetřením mezi žáky spojen s žádnými významnými důsledky pro samotné učitele. Pozorování ve třídě byla navíc hodnocena externími hodnotiteli. Vzhledem k tomu, že federální politiky tlačí státy a školské distrikty, aby rozšířily hodnocení a zvažovaly i měření výkonu nezaložená na testech, bude pro budoucí studie důležité prozkoumat prediktivní validitu měření vlivů učitelů na výsledky žáků v různých prostředích, v různých mírách odpovědnosti učitelů za výsledky žáků a s různými typy hodnotitelů (jako jsou škola, administrátoři ze školského distriktu...).

3.6 Literatura

Bacher-Hicks, A., Chin, M. J., Kane, T. J., Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, 73, 101919.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (No. w20657). National Bureau of Economic Research.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, 13(3), 281–309.

Conover, W. J. (1999). *Practical nonparametric statistics* (Vol. 350). John Wiley & Sons. Hoboken, NJ.

Ferguson, R. (2009). Tripod student survey, MET project upper elementary and MET project secondary versions. Westwood, MA: Cambridge Education.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, 37(2), 224–242.

Glazerman, S., & Protik, A. (2015). Validating value-added measures of teacher performance. *Association for Public Policy Analysis & Management*, November, Miami.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job* (pp. 2006–01). Washington, DC: Brookings Institution.

Hickman, J. J., Fu, J., & Hill, H. C. (2012). Creation and dissemination of upper-elementary mathematics assessment modules. Princeton, NJ: ETS.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathe-

mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, 26(4), 430–511.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the impacts of teachers: Reply. *American Economic Review*, 107(6), 1685–1717.

Jacob, B., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education*. National Bureau of Economic Research, Cambridge, MA (2005). (No. w11463).

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Research Paper. MET Project. Bill & Melinda Gates Foundation*.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education review*, 27(6), 615–631.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education review*, 27(6), 615–631.

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education*, 123(4), 615–646.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, 29(1), 67–101.

Minnici, A. (2014). The Mind Shift in Teacher Evaluation: Where We Stand--and Where We Need to Go. *American Educator*, 38(1), 22–26.

National Governors Association Center for Best Practices & Council of Chief State School Officers. Common Core State Standards Initiative. (2010). Common core state standards for mathematics. http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.

Papay, J. P., Bacher-Hicks, A., Page, L. C., & Marinell, W. H. (2017). The challenge of teacher retention in urban schools: Evidence of variation from a cross-site analysis. *Educational Researcher*, 46(8), 434–448.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review*, 94(2), 247–252.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review*, 107(6), 1656–1684.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis*, 38(2), 293–317.

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating teachers with classroom observations. *Brown Center on Education Policy: Brookings Institute*.

4. Hodnocení výkonu učitelů prostřednictvím jejich přidané hodnoty v praxi škol³³

District of Columbia (Washington, D. C.) byl jedním z prvních školských obvodů ve Spojených státech, který se pokusil propojit individuální přístup k hodnocení výkonu učitele s důležitými rozhodnutími o jeho platu a ochotě jej zaměstnávat. Ve školním roce 2009/2010 byl v D. C. spuštěn nový systém hybridního hodnocení učitelů – tzv. projekt IMPACT. V rámci tohoto projektu byly explicitně formulovány standardy pro poskytování efektivní výuky, učitelům byli přidělováni koučové, kteří jim pomáhali tyto standardy plnit a poskytovali jim zpětnou vazbu. Zároveň všichni učitelé podléhali pravidelnému každoročnímu hodnocení. Na konci každého školního roku bylo učitelům přidělováno výsledné skóre z proběhlých hodnocení, které se nacházelo v rozmezí 100 až 400 bodů. Učitelé byli hodnoceni na základě jejich výkonu ve třech klíčových oblastech: (1) posouzení kvality jimi poskytované výuky (informace o kvalitě výuky byly získávány z hospitací), (2) výkon žáků ve standardizovaných testech a (3) oddanost školní komunitě. Jednotlivé komponenty výsledného skóre a jejich váhy v celkovém hodnocení se ovšem lišily v závislosti na tom, jaká konkrétní měření výkonu byla u daných učitelů k dispozici (ne všechny ukazatele výkonu učitelů totiž byly k dispozici u všech učitelů). Níže jsou popisovány složky systému hodnocení tak, jak existovaly v prvních třech letech po zavedení programu do praxe, protože v pozdějších obdobích docházelo k jejich úpravám.

Ve zmiňovaném období bylo jednou z klíčových komponent hodnocení učitelů hodnocení jejich výkonu během hospitací. U všech učitelů bylo realizováno celkem pět strukturovaných hospitací (pozorování a strukturované hodnocení kvality výuky). Hodnocení kvality výuky při hospitacích probíhalo v souladu s pokyny uvedenými v *Teaching and Learning Framework*, což je rámec, který definuje domény efektivní výuky, např. používání žádoucích organizačních forem výuky; zaměření výuky na cíl; kontrolu porozumění probírané látce; jasné vysvětlení obsahu; efektivní využívání času. Z pěti předepsaných pozorování byla tři pozorování realizována řediteli škol a dvě zkušenými pedagogy pověřenými hospitacemi na několika školách v dané lokalitě. Učitelé byli dopředu seznámeni pouze s termínem první hospitace. Váha skóre získaného z hospitací se lišila v závislosti na tom, zda byl u daného učitele k dispozici i odhad jeho přidané hodnoty, tedy skóre, které se odvíjelo od výkonu jeho žáků ve standardizovaných testech. Tento ukazatel byl k dispozici zejména u učitelů čtení a matematiky ve 4. – 8. ročnících, kteří ovšem tvořili pouze méně než jednu pětinu všech učitelů. V případě těchto učitelů představovaly výsledky z hospitací pouze 35 procent celkového skóre, protože do jejich celkového výsledku se započítávala i jejich modelem odhadnutá přidaná hodnota, která činila 50 procent z celkového skóre. Pro většinu ostatních učitelů, pro které nebyly modelové odhady jejich přidané hodnoty k dispozici (např. z důvodu, že žáci v jimi vyučovaných předmětech nepodstupovali standardizované testy), jejich skóre z hospitací činilo 75 procent celkového skóre. Další 10 procent z celkového skóre bylo u těchto učitelů postaveno na hodnocení jejich pokroku směrem k cílům, které si každoročně stanovovali

³³ Tento text je podrobným výtahem z článku: Dee, T., Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. *Education Next*, 17(4), 58–67. Z tohoto textu je přebírán i odkazový aparát.

ve spolupráci s řediteli svých škol. Následně dalších 10 procent z celkového skóre hodnocení učitelů činil jejich příspěvek učící se komunitě školy, což byla charakteristika, kterou posuzoval ředitel školy. Posledních 5 procent z bodového ohodnocení bylo založeno na míře přidané hodnoty učitele k výsledkům žáků školy jako celku. Učitelům mohly být body i odečteny, a to za problémy s docházkou, nedochvilnost a nedodržování předepsaných pravidel a postupů.

4.1 Rozdíly mezi učiteli

Na rozdíl od typických systémů hodnocení učitelů, nový systém hodnocení vedl k velmi podstatné diferenciaci v hodnocení učitelů. Ve školních letech 2009/2010 a 2010/2011 bylo v rámci nového systému hodnocení 14 procent učitelů hodnoceno jako „*vysoce efektivní*“, 69 procent jako „*efektivní*“, 14 procent jako „*minimálně efektivní*“ a další 2 procenta byla považována za „*neefektivní*“³⁴.

Nový systém hodnocení zároveň předepisoval, jaké konkrétní dopady má mít na učitele nízké nebo naopak velmi vysoké hodnocení. Učitelé, kteří byli hodnoceni jako „*neefektivní*“ byli až na vzácné výjimky propouštěni, což způsobilo, že v prvním roce po zavedení systému opustilo školy v D. C. více než 200 učitelů. Ti, kteří byli hodnoceni jako „*minimálně efektivní*“, měli jeden rok na to, aby zvýšili svoje skóre, a pokud tak neučinili, byli propuštěni. Na druhé straně spektra byli učitelé, kteří byli hodnoceni jako „*vysoce efektivní*“, ti měli nárok na roční bonus až 25 000 USD. Těm z nich, kteří byli hodnoceni jako „*vysoce efektivní*“ dva nebo více školních roků po sobě, mohla být zvýšená základní mzda až o 27 000 USD ročně. Také učitelé pracující ve školách s vysokou mírou chudoby a učitelé vyučující předměty vysoké důležitosti měli nárok na zvýšení platů. Tyto cílené strategie vytvořily ostré motivační kontrasty pro učitele nacházející se na obou stranách prahu odlišující „*efektivní*“ učitele od učitelů „*minimálně efektivních*“. Skórování nad tímto prahem minimalizovalo hrozbu propuštění a zároveň pro „*vysoce efektivní*“ vytvářelo prostor pro výrazné zvýšení platu, a naopak, „*neefektivním*“ a „*minimálně efektivním*“ učitelům reálně hrozilo propuštění.

4.2 Cíle studie

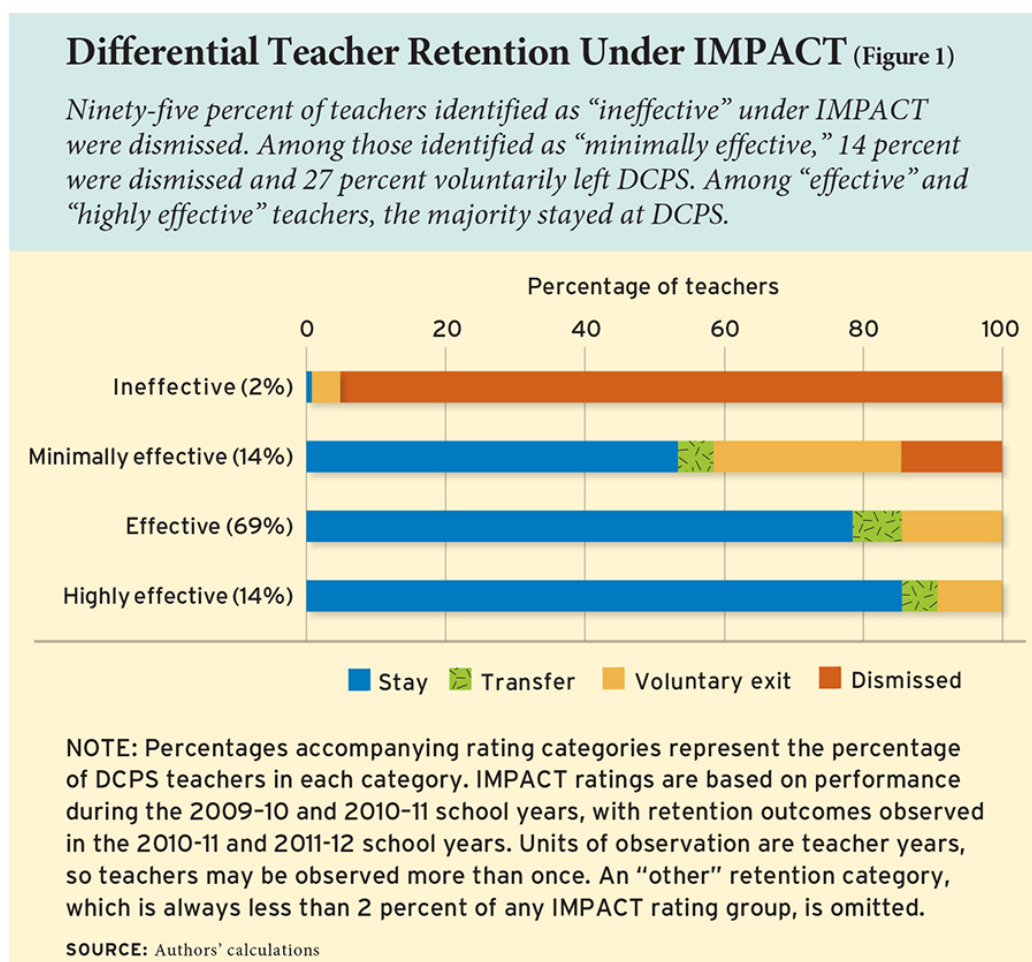
Cílem autorů této studie bylo pochopit, zda tyto přístupy k hodnocení učitelů ovlivňují výkon učitelů a jejich udržení ve škole, a zda zlepšují výsledky žáků. Analýza byla založena na administrativních údajích o všech učitelích v D. C. a jejich žácích v prvních třech letech po zavedení nového systému hodnocení, tedy ve školních rocích: 2009/2010, 2010/2011 a 2011/2012. Vzorek byl omezen na učitele všeobecně vzdělávacích předmětů pracující ve školách K–12. Při hodnocení vlivu nového systému hodnocení na výsledky žáků v testovaných ročnících a předmětech bylo čerpáno z údajů získávaných ještě v dalším školním roce – 2012/2013.

³⁴ Kvůli zaokrouhlování na celá čísla není jejich součet 100 procent.

4.3 Udržení a výkon učitele

Dle nového systému hodnocení závisela snaha školy o udržení učitele na jeho pozici na hodnocení jeho výkonu. Během prvních tří let od zavedení nového systému bylo ze škol propuštěno 95 procent učitelů z těch, kteří byli v prvních dvou letech hodnoceni jako „neefektivní“ (viz obrázek 1). Celkově bylo propuštěno 3,8 procent všech učitelů v důsledku toho, že byli v letech 2009/2010 až 2011/2012 jednou hodnoceni jako „neefektivní“ nebo dvakrát po sobě hodnoceni jako „minimálně efektivní“. Zároveň nový systém hodnocení způsobil, že u učitelů s nižším hodnocením byla mnohem větší pravděpodobnost, že ze škol odejdou dobrovolně sami. 27 procent ze všech učitelů, kteří byli hodnoceni jako „minimálně efektivní“, dobrovolně opustilo školy, v porovnání se 14 procenty učitelů, kteří byli hodnoceni jako „efektivní“ a 9 procenty učitelů, kteří byli hodnoceni jako „vysoce efektivní“.

Obrázek 1: Různá míra udržení učitelů v rámci nového systému hodnocení



Poznámka: Převzato z: Dee, T., Wyckoff, J. (2017). *A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. Education Next*, 17(4), 58–67.

„*Minimálně efektivní*“ učitelé, jejichž skóre se nejvíce blížilo prahu odlišujícího „*minimálně efektivní*“ učitele od učitelů „*efektivních*“, měli menší pravděpodobnost odchodu než učitelé s nižším skóre. Přibližně každý čtvrtý učitel, jehož skóre bylo v rozmezí 25 bodů od „*efektivní*“ prahové hodnoty, se rozhodl odejít ze zaměstnání, v porovnání s přibližně jedním ze tří, jehož skóre bylo o více než 25 bodů nižší. Jinými slovy, učitelé pod hrozbou propuštění častěji dobrovolně odcházeli než učitelé, kteří této hrozbě nepodléhali, a ti, kteří měli nejdále od „*efektivní*“ prahové hodnoty, měli ještě větší pravděpodobnost, že odejdou.

Tyto vzorce chování učitelů byly v souladu se zamýšlenou restrukturalizací učitelské pracovní síly, ale existují pro ně i jiná vysvětlení. Některé studie například zjistily, že v rané fázi kariéry z profese odcházejí méně efektivní učitelé s vyšší pravděpodobností než efektivnější začínající učitelé, a to i při absenci systémů hodnocení. Navíc skóre v rámci nového systému hodnocení bylo pro začínající učitele v prvních dvou letech jejich pedagogické praxe v průměru o 17 bodů nižší než u učitelů s tříletou a delší praxí. Takové úvahy vyvolávají pochybnosti o tom, jak interpretovat vztah mezi novým systémem hodnocení a udržení učitelů ve školách. Otázkou je, zda ve výše zmíněných vzorcích chování pozorujeme spíše účinky přístupů (pobídek) daných novým systémem hodnocení, nebo chování, ke kterému by došlo i bez přijetí tohoto způsobu hodnocení.

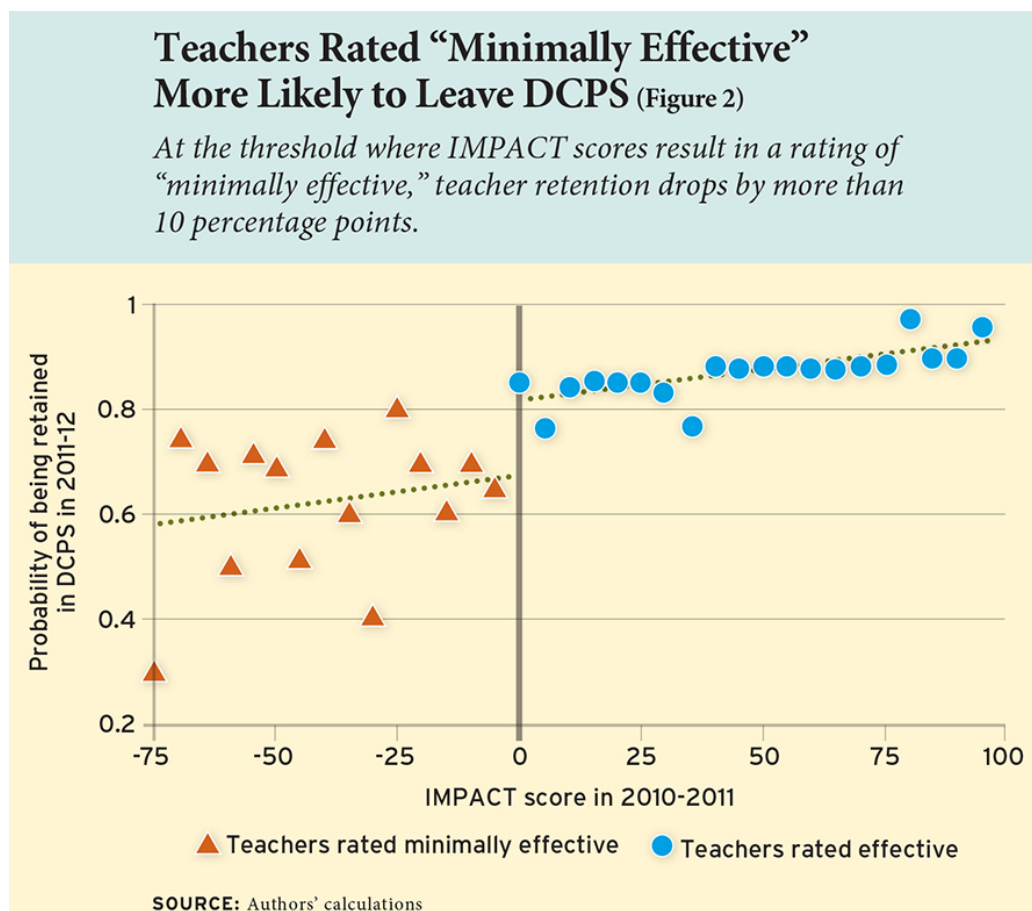
Aby autoři studie mohli izolovat kauzální vliv pobídek založených na výkonu na kvalitu učitelů, porovnávali výsledky mezi učiteli, jejichž počáteční skóre je umístilo blízko prahových hodnot mezi kategoriemi: „*minimálně efektivní / efektivní*“ a „*efektivní / vysoce efektivní*“. Předpokládali, že to, zda učitel dosáhne skóre těsně nad nebo těsně pod určitou prahovou hodnotou, je v podstatě náhodné, což je podpořeno jak znalostmi o tom, jak distrikt vypočítává skóre, tak dalšími analýzami potvrzujícími, že učitelé na obou stranách každé prahové hodnoty jsou si z hlediska zkušenosti a další vlastností velmi podobní. Autoři studie sledovali schopnost učitele udržet se ve škole v průběhu druhého roku projektu, kdy se hrozba propuštění pro „*minimálně efektivní*“ učitele stala nově velmi reálnou a finanční pobídky pro „*vysoce efektivní*“ učitele mohly být trvalé. Zkoumali také výkon u vracejících se učitelů ve třetím roce fungování systému hodnocení. Porovnáním dobrovolných odchodů učitelů a výkonu učitelů na každé straně spektra získali lepší představu o tom, jak hrozba propuštění nebo vyhlídka na zvýšení platu ovlivňuje chování učitelů.

Tuto metodu autoři studie poprvé použili k měření toho, jaký efekt měl nový způsob hodnocení učitelů na učitele, kteří skórovali v blízkosti prahu „*minimálně efektivní / efektivní*“ v letech 2010/2011. Zjistili, že učitelé poblíž prahu, kteří v této době obdrželi své první hodnocení „*minimálně efektivní*“, měli podstatně větší pravděpodobnost, že školu dobrovolně opustí, přičemž jejich retence klesla v letech 2011/2012 o deset procentních bodů (viz obrázek 2). Jinými slovy, u minimálně efektivních učitelů se zvýšila pravděpodobnost jejich odchodu z dvaceti procentních bodů na třicet procentních bodů, což je nárůst o 50 procent.

Mezitím učitelé v této „*minimálně efektivní*“ skupině, kteří zůstali na pozici učitele i v následujícím školním roce, zlepšili svůj výkon zhruba o 11 bodů. Tento prudký nárůst výkonu učitelů činí asi 0,27 směrodatné odchylky (viz obrázek 3) a naznačuje, že dříve slabí učitelé, kteří se rozhodli dále působit jako učitelé, se úspěšně zlepšili.

Je pozoruhodné, že účinky hodnocení „*minimálně efektivní*“ na udržení učitele v profesi i jeho výkon se objevily až na konci druhého roku po zavedení systému hodnocení, kdy byla politická důvěryhodnost refor-

Obrázek 2: Učitel, který byl hodnocen „minimálně efektivní“ s vyšší pravděpodobností opustil distrikt



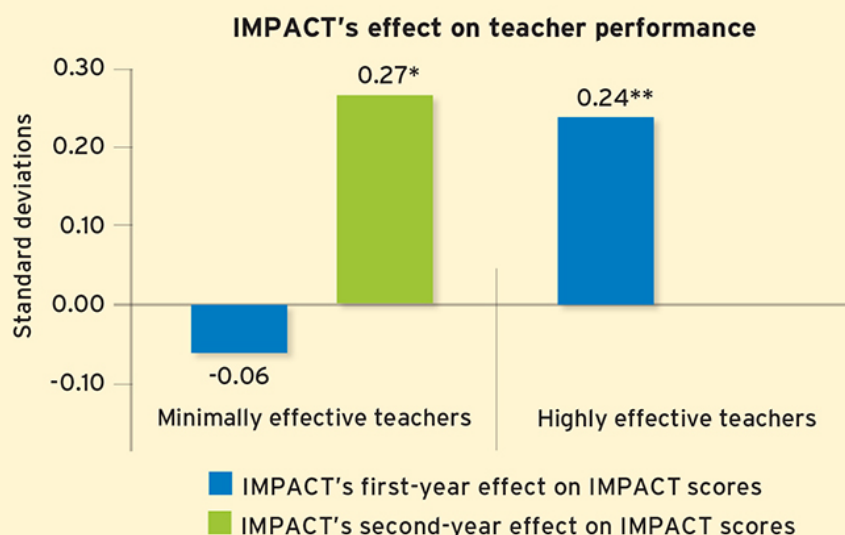
Poznámka: Převzato z: Dee, T., Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. *Education Next*, 17(4), 58–67.

my potvrzena jmenováním pedagožky a podporovatelky reforem Kayi Henderson do pozice kancléřky zodpovědné za vzdělávání na veřejných školách v DC, a poté co zhruba 140 učitelů bylo propuštěno, protože měli ve dvou po sobě jdoucích hodnoceních výsledek „minimálně efektivní“. Po prvním roce implementace nového systému hodnocení nebyl identifikován žádný statisticky zjiitelný vliv na udržení učitelů ve škole nebo na jejich výkon. Vzhledem ke kontroverznosti těchto politik a politické volatilitě spojené s odchodem starosty Adriana Malik Fentyho a rezignací předchozí kancléřky zodpovědné za školství Michelle Ann Rhee, mohli učitelé důvodně pochybovat o reálnosti hrozeb propouštění na základě výkonu a mohli se rozhodnout nereagovat na pobídky.

Důležité ponaučení je, že politiky, které mají za cíl přimět aktéry k významné změně chování, mohou vyžadovat čas, aby se uchytily. Naproti tomu finanční pobídky pro trvale vysoce výkonné učitele, které znamenaly skok v jejich platovém ohodnocení, se zdály být učiteli okamžitě akceptované.

Obrázek 3: Zvýšení výkonu učitelů**Teacher Performance Boost (Figure 3)**

A “minimally effective” rating in IMPACT’s first year had small and statistically insignificant effects on subsequent teacher performance (relative to an “effective” rating). However, in IMPACT’s second year, teachers rated “minimally effective” who chose to remain in DCPS improved their performance by 0.27 standard deviations compared teachers rated “effective.” Meanwhile, observed in IMPACT’s first year, the financial incentives available to teachers on the “highly effective” side of the threshold improved subsequent teacher performance by 0.24 standard deviations.



* Statistically significant at the 95% confidence level

** Statistically significant at the 99% confidence level

NOTE: First-year effects are the 2009-10 IMPACT rating effects on outcomes observed in 2010-11, and second-year effects are the 2010-11 IMPACT rating effects on outcomes observed in 2011-12. This analysis includes only the first-year effects to assess the effect of receiving a rating of “highly effective” for the first time, which triggered a bonus of up to \$25,000. Including teachers rated “highly effective” in the second year could also include teachers receiving a permanent salary boost triggered by repeated high ratings. However, including data from IMPACT’s second year leave our results qualitatively unchanged.

SOURCE: Authors’ calculations

Poznámka: Převzato z: Dee, T., Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. *Education Next*, 17(4), 58–67..

Retence učitelů, kteří byli konzistentně hodnoceni jako vysoce efektivní a dostali se velmi blízko tomu, aby byli způsobilí pro trvalé zvýšení platu, se zvýšila zhruba o tři procentní body (i když tento efekt nebyl statisticky významný) a jejich výkon se v letech 2011/2012 zlepšil o statisticky významných 10,9 procentních bodů neboli 0,24 směrodatné odchylky. Tyto nárůsty výkonu znamenají zvýšení přibližně o pět procentních bodů v hodnocení výkonu u učitelů s nižším hodnocením a 7 procentních bodů u učitelů s vysokým hodnocením. Porovnáme-li tato čísla s typickým zlepšením začínajícího učitele během prvních tří let ve třídě, zjistíme, že zisk 12,6 bodu u učitelů s nižším hodnocením představuje 52 procent tohoto běžně pozorovaného tříletého zisku a zisk 10,9 bodu vysoce hodnocených učitelů je 41 procent tohoto tříletého zisku.

4.4 Fluktuační učitelů a výsledky žáků

Výsledný efekt nového systému hodnocení mimo jiné závisí na přímých důsledcích dopadu fluktuace učitelů na žáky a na kvalitě nově přijatých učitelů. Často se předpokládá, že fluktuace učitelů negativně ovlivňuje kvalitu škol a vysoká míra fluktuace komplikuje náborové procesy. Nicméně ne všechna fluktuace je stejná, a ne všechny dopady fluktuace jsou totožné. Např. v rámci nového systému hodnocení se podstatná část fluktuace týkala méně výkonných učitelů, kteří byli záměrně povzbuzováni k odchodu ze školy, aby tímto způsobem školy pozitivně ovlivnily distribuci efektivit stávajících učitelů

Autoři textu se zaměřili na zkoumání toho, jaký vliv měla fluktuace, ke které došlo v rámci nového systému hodnocení, na výsledky žáků. Zkoumali meziroční změny ve třídách s fluktuací učitelů a bez ní. Jinými slovy, zjišťovali, jaká je meziroční změna ve výsledcích žáků 4. tříd ve standardizovaných testech ve škole, které měla v daném ročníku fluktuující učitele, v porovnání se změnou skóre v testech žáků ve škole, která v tomto ročníku neměla fluktuaci učitelů.

Zjistili, že celkový efekt fluktuace ve školách v D. C. neměl žádný negativní dopad na výsledky žáků a za rozumných okolností fluktuace výsledky žáků dokonce zlepšila. Tento celkový efekt fluktuace kombinuje negativní efekt³⁵ odchodu vysoce výkonných učitelů s efektem odchodu učitelů s nízkými výkony. Obrázek 4 porovnává průměrné skóre přidávané hodnoty nových učitelů, kteří vstoupili do distriktu, s těmi, kteří odešli v předchozím roce.

Při sledování dopadů na výsledky žáků a po pečlivé kontrole vlivu různých dalších faktorů, které mohou ovlivnit výsledky (tzv. *confounding factors*), došli autoři studie k závěru, že fluktuace učitelů neměla významný negativní dopad na výsledky žáků (viz obrázek 5).

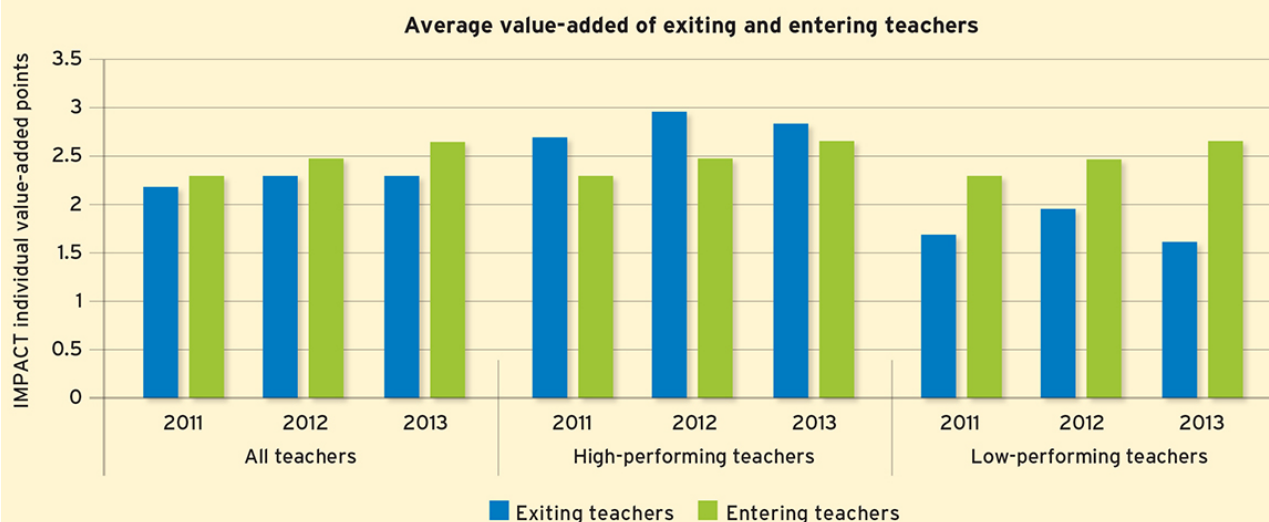
Z grafů uvedených v obrázku 5 vyplývá, že fluktuace učitelů s nízkým výkonem vede k lepším výsledkům žáků v matematice i ve čtení. Například v matematice odchod učitelů s nízkou výkonností podle odhadů zlepšil výsledky žáků o 0,21 směrodatné odchylky, tj. rozdíl, který odpovídá třetině až dvou třetinám roku učení v závislosti na úrovni ročníku. Ve čtení se výsledky žáků zlepšily o 0,14 standardní odchylky.

Autoři studie porovnávali také rozdíly ve fluktuaci a jejich dopad na výsledky žáků ve školách s vysokou a nízkou mírou chudoby. Podle nich se zdá říci, že školy s vysokou mírou chudoby se v důsledku fluktuace učitelů

³⁵ Tento negativní efekt ovšem nebyl statisticky významný.

Obrázek 4: Sledování kvality učitelů**Tracking Teacher Quality (Figure 4)**

Teacher turnover under IMPACT improved teacher quality. Although high-performing teachers who left DCPS were replaced by somewhat less-effective teachers, on average, when low-performing teachers left DCPS, they were replaced by much more effective teachers, on average.



NOTE: Results for 2011, for example, indicate the average score for all teachers who exited at the end of 2009-2010 compared with all those entering in 2010-2011. Exits include teachers who retired, resigned, or were terminated. Teachers leaving schools that closed are excluded. Average teacher value-added was converted to a 0-4 scale using a conversion table. Data are for teachers who are matched to students with math achievement scores.

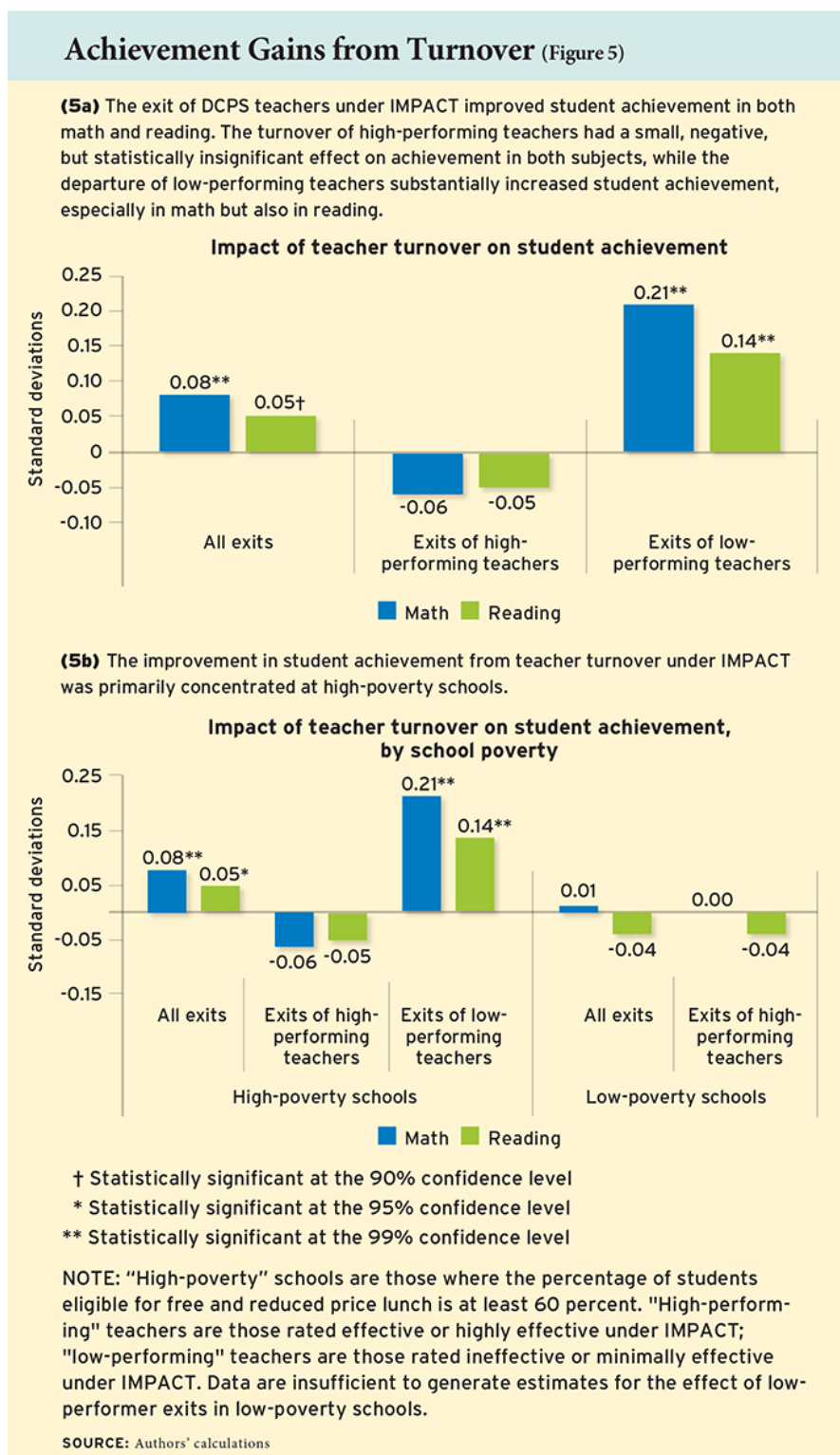
SOURCE: Authors' calculations

Poznámka: Převzato z: Dee, T., Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. Education Next, 17(4), 58–67.

zlepšují, i když ve všech školách není fluktuace stejná. Ve školách s vysokou mírou chudoby je odhadováno, že celkový vliv fluktuace na výsledky žáků je 0,08 směrodatné odchylky v matematice a 0,05 směrodatné odchylky ve čtení. Naproti tomu, ve školách s nízkou mírou chudoby se odhadované dopady fluktuace blíží nule.

Nicméně 40 procent fluktuace učitelů ve školách s vysokou mírou chudoby je mezi učiteli s nízkou výkonností. Odhady autorů naznačují, že lze identifikovat trvale velké zisky pro žáky, když tito učitelé tyto školy opustí. Například v matematice se výsledky žáků zlepšily o 0,21 standardní odchylky, přičemž kvalita učitele se zlepšila o 1,3 standardní odchylky. Ve čtení se výsledky žáků zlepšily o 0,14 směrodatné odchylky a kvalita učitelů se zlepšila o 1,0 směrodatnou odchylku. Stručně řečeno, ve veřejných školách v D. C. se podařilo v rámci nového způsobu hodnocení nahradit učitele s nízkou výkonností ve školách s vysokou mírou chudoby jinými učiteli, kteří jsou podstatně efektivnější. Zdá se, že systém byl i schopný nahradit odcházející vysoce výkonné učitele ve školách s nízkou mírou chudoby srovnatelnými učiteli. S výjimkou prospěchu v matematice a čtení v jednom školním roce (2011/2012) však vliv této fluktuace na prospěch žáků není statisticky významný.

Obrázek 5: Zisky ve výsledcích žáků jako důsledek fluktuace



Poznámka: Převzato z: Dee, T., Wyckoff, J. (2017). *A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. Education Next*, 17(4), 58–67.

4.5 Možnosti implikace

Zavedení nového systému hodnocení učitelů bylo zdrojem sporů, a to jak v rámci D. C., tak i v širších diskusích o vzdělávací politice. Zatímco důležitost vysoce efektivních učitelů je nekontroverzní, způsob, jakým školy a okresy tyto pedagogy identifikují, odměňují a udržují, je předmětem diskuze a kontroverze. Tato studie poskytuje jasný důkaz, že silné, individuálně cílené pobídky spojené s výkonem ovlivňují udržení učitelů v profesi a jejich výkon žádoucím způsobem. Tento systém vícenásobných opatření zvyšuje výkon učitelů a podporuje vyšší míru fluktuace mezi učiteli s nehoršími výsledky, což má pozitivní dopady na výsledky žáků.

Důležité je, že více než 90 procent fluktuace učitelů s nízkými výsledky v hodnocení se vyskytuje ve školách s vysokou mírou chudoby, které tvoří 75 procent všech škol v D. C. Podíl odcházejících učitelů se špatnými výsledky je v těchto školách s vysokou mírou chudoby dvakrát vyšší, než ve školách s nízkou mírou chudoby. V porovnání s téměř jakýmkoli jiným zásahem jde o velmi velká zlepšení pro některé z nejpotřebnějších žáků. I když zavedení nových systémů hodnocení vedlo k propuštění některých učitelů, byla patrná i určitá míra dobrovolné fluktuace.

Autoři udávají, že by tvůrci politik, kteří uvažují o zavedení obdobných systémů hodnocení, měli vědět, že vždy existuje riziko určitého počtu nežádoucích chyb při hodnocení a vyvozování důsledků z hodnocení. Minimalizace těchto chyb bude vyžadovat drahé a sofistikované nástroje. Tvůrci politik musí tyto náklady porovnat se vzdělávacími a ekonomickými přínosy, které tyto systémy mohou mít pro žáky, a to jak tím, že se zamezí udržování učitelů s nehoršími výsledky v jejich pracovních pozicích, tak i prostřednictvím významného zvýšení výkonu učitelů a pracovní síly obecně.

Mělo by se také počítat se změnami a vylepšováním různých systémů hodnocení. D. C. téměř každý rok prováděl úpravy systému hodnocení, například snižoval počet hospitací učitelů, měnil přístup k finančním benefitům i ke zvýšení základního platu ve školách s vysokou mírou chudoby a snižoval váhu měření výkonu postaveného na měření přidané hodnoty učitelů opírající se o výsledky testů žáků. Nedávno D. C. oznámil, že se bude při všech hospitacích spoléhat na ředitele škol a začlení do hodnocení výsledky hodnocení učitelů žáky získaných prostřednictvím žákovských dotazníků. Rovněž nově představil program intenzivního profesního rozvoje LEAP, který má učitelům pomoci zlepšit jejich dovednosti. Stále trvá snaha zlepšit složení pracovní síly směrem k vyšší efektivitě učitelů a škol.

4.6 Literatura

Dee, T., Wyckoff, J. (2017). A lasting impact: High-stakes teacher evaluations drive student success in Washington, DC. *Education Next*, 17(4), 58–67.

