

Vztah rychlého odpovídání v testu z anglického jazyka a výsledného skóre u různých skupin žáků

Jiří Štípek, Hana Voňková, Kateřina Králová

Univerzita Karlova, Pedagogická fakulta

Abstrakt: Studie se zaměřuje na úsilí žáků při odpovídání na testové otázky a jeho vztah k výslednému skóre v testu. Cílem studie bylo zjistit, v jaké míře se nedostatečné úsilí při testování vyskytuje u různých skupin žáků dle genderu a typu školy a jak se rozdílná míra nedostatečného úsilí u těchto skupin žáků může promítat do jejich srovnávání na základě výsledného testového skóre. Výzkumný vzorek tvořili žáci a žákyně 9. ročníků základních škol ($N = 631$) a odpovídajících ročníků víceletých gymnázií ($N = 713$). Výzkumným nástrojem byla on-line dotazníková aplikace, jejíž součástí byl rozřazovací test z anglického jazyka (Cambridge EFL Placement Test) s 25 uzavřenými výběrovými otázkami (multiple-choice item). K identifikaci nedostatečného úsilí jsou užívány časové charakteristiky chování žáků během testování, které zaznamenávala on-line testovací aplikace vyvinutá za tímto účelem. Výsledky studie ukazují, že u žáků víceletých gymnázií je nedostatečné úsilí okrajovým jevem (3,1 %), zatímco u žáků základních škol se týká téměř jedné čtvrtiny (23,6 %). V obou typech škol přitom chlapci vykazují nedostatečné úsilí přibližně dvakrát častěji než dívky. Analýza výsledků v testu naznačuje, že nižší skóre u chlapců ve srovnání s dívkami může být z větší části vysvětleno právě nižším úsilím chlapců při vyplňování testu. Naše zjištění naznačují, že nedostatečné úsilí při vyplňování testů může mít významný vliv na závěry studií, které na základě výsledků testů porovnávají různé skupiny žáků, především pokud jde o testy bez konkrétních důsledků pro samotného žáka (tzv. low-stakes tests).

Klíčová slova: didaktický test, úsilí žáků, čas na odpověď, testové skóre, rapid guessing behavior

The Relationship Between Fast Response Time in English Language Test and The Final Score Among Different Groups of Students

Abstract: This study focuses on the effort of students when answering achievement test questions and its relationship with their final test score. The aim of the study was to determine the extent to which insufficient effort responding in tests occurs among different groups of students based on gender and school type, and how the varying degrees of insufficient effort among these groups may be reflected in their comparison based on the final test score. The research sample consisted of male and female students from the 9th grade of basic school ($N = 631$) and the corresponding grades of multi-year secondary general schools ($N = 713$). As a research tool, we employed an online questionnaire application, which included an English placement test (Cambridge EFL Placement Test) with 25 multiple-choice questions. To identify insufficient effort, we employ time characteristics of students' behavior during testing, recorded by an online testing application developed for this purpose. The results of the study indicate that insufficient effort is a marginal phenomenon among students of multi-year secondary general schools (3.1%), while affecting almost a quarter of basic school students (23.6%). In both school types, boys exhibit insufficient effort approximately twice as much as girls. Analysis of test results suggests that lower scores of boys compared to girls may be largely explained by lower effort of boys when completing the test. Our findings suggest that insufficient effort responding in tests may have a significant impact on conclusions of studies that compare different groups of students based on test results.

Keywords: achievement test, students' effort, response time, test score, rapid guessing behavior

<https://doi.org/10.14712/23363177.2025.7>
www.orbisscholae.cz

© 2025 The Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>).

42 Součástí vzdělávání ve školách je také zjišťování jeho výsledků pomocí různých testů. Přestože testy mohou poskytovat poměrně přesnou informaci o výsledcích vzdělávání, existují vedle žákovy úrovně měřených konceptů i další faktory, které vstupují do procesu testování a mohou ovlivnit výsledné testové skóre. Jedním z nich je míra úsilí, které žák během testování vynakládá. Wise a DeMars (2005) ve své přehledové studii ukazují, že testovaní, kteří k testu přistupují s nedostatečným úsilím, dosahují o více než polovinu směrodatné odchylky nižší skóre v testu než ostatní. Z toho vyplývá, že se uvedený jev promítá do výsledků testu, což může následně i snížit platnost závěrů učiněných na jejich základě (např. Meade & Craig, 2012; Wise & Gao, 2017).

Pro zjišťování přístupu žáků k testování byla vyvinuta řada prostředků, např. kontrolní otázky, obrácené škály, či analýza řad odpovědí (více např. Curran, 2016). S nástupem informačních technologií, resp. využíváním elektronických testů se však objevily nové možnosti, jak některé projevy vynaloženého úsilí zachytit. Soudobé testovací nástroje, využívající digitální technologie, mohou přesně evidovat řadu aktivit testovaného, které provádí během testování (např. Wise & Gao, 2017). Vedle odpovědí na otázky lze zaznamenat například pořadí řešení otázek, změny odpovědí či potvrzení odpovědi. Evidence aktivit může být dostupná pro každou položku testu, či dokonce její část (např. konkrétní distraktor). Zaznamenávat lze jak samotné akce, tak i čas jejich provedení. Získaná data pak poskytují poměrně podrobné časové charakteristiky interakce testovaného s testem. Na jejich základě pak můžeme usuzovat na chování a změny chování každého testovaného v průběhu celého testování a vyvozovat z něj, zda vynaložil dostatečné úsilí. Dostupnost takových údajů (tzv. log data) vedla v posledních dvaceti letech ke vzniku řady nových metod identifikace nedostatečného úsilí (přehledy např. Guo et al., 2016; Kröhne et al., 2020; Sahin & Colvin, 2020; Wise, 2019), založených především právě na analýze zmíněných časových charakteristik.

Nově vzniklé metody vycházejí z předpokladu, že zodpovězení každé položky (otázky) testu vyžaduje určitý minimální čas. Tento minimální čas je chápán jako nejkratší doba nutná k přečtení položky (otázky, nabízených možností odpovědi atp.), pochopení položky, zvážení odpovědi a samotného aktu odpovědi. Časy kratší, než minimální jsou chápány jako projev nedostatečného úsilí (Wise & Kong, 2005), jež může mít různé primární příčiny (nedostatečná motivace, únava apod.), a v literatuře je pro takové chování používána řada termínů. Nejčastěji se objevují *rapid guessing behavior* a *speeding*, které však v zásadě znamenají totéž a odkazují k chování, kdy respondent na položku odpoví v tak krátkém čase, že ji nemohl stihnout přečíst a plně zvážit (Kong et al., 2007; Lee & Jia, 2014; Schnipke & Scrams, 1997). V praxi se však termín *rapid guessing behavior* objevuje zejména ve studiích zabývajících se testy a termín *speeding* spíše v souvislosti s dotazníky.

V této studii bude proto dále používán termín *rapid guessing behavior*, resp. *rapid guessing*, a navíc též komplementární termín *solution behavior* (Schnipke & Scrams, 1997; Wise, 2017), který označuje aktivní snahu respondenta správně odpovědět na testové položky.

Jako podstatné z hlediska vynaloženého úsilí se jeví i to, jaký význam má úspěšné složení testu pro testované. Literatura odlišuje tzv. high-stakes a low-stakes testy, resp. testování. U *high-stakes* testování má výsledek testu důsledky pro testované, např. má vliv na hodnocení v daném předmětu nebo se promítá do výsledku přijímacího řízení. V případě *low-stakes* testování jsou důsledky pro testované minimální nebo zcela absentují. Studie, které se zabývají high-stakes a low-stakes testováním ukazují, že absence důsledků u low-stakes testování vede k větší míře nedostatečného úsilí, což je interpretováno jako doklad nižší motivace testovaných (Duckworth et al., 2011; Wise & DeMars, 2005).

Dosavadní aplikace nových metod ve výzkumu ukazují, že nedostatečné úsilí není okrajovým jevem, ale zjištěná míra tohoto chování napříč studiemi je poměrně variabilní a záleží jak na testované oblasti, tak i na skupině testovaných a dalších faktorech. Rios et al. (2022) ve své metaanalýze studií založených na analýze časových charakteristik testů zjistili, že v průměru 7 % odpovědí je identifikováno jako projev nedostatečného úsilí a že k takovému chování se alespoň jednou uchýlí v průměru 28 % testovaných. V případě low-stakes testování se však lze setkat i s tím, že procento odpovědí identifikovaných jako projev nedostatečného úsilí přesahuje 10 % (Welling et al., 2024) a u jednotlivých otázek pak i čtvrtinu odpovědí (Goldhammer et al., 2017). Wise et al. (2009) ve své studii zjistili, že během testu, který měl 64 otázek s výběrovou odpovědí (multiple-choice item), administrovaného za low-stakes podmínek, 47 % studentů projevilo nedostatečné úsilí alespoň jednou a 24 % studentů se k takovému chování uchýlilo u více než 10 % otázek (7 a více).

Výzkumy dále přinášejí zjištění, že existují rozdíly mezi různými podskupinami respondentů, např. u chlapců je podíl jedinců vykazujících nedostatečné úsilí větší než u dívek (DeMars et al., 2013; Soland, 2018). V šetření PISA 2015 byla podle zjištění Anayaové a Zamarrové (2024), které ve své studii analyzovaly data 55 států¹, míra nedostatečného úsilí u dívek 2,8 % a u chlapců 3,8 %. Rozdíly jsou i mezi různými stupni škol (Kornhauser et al., 2014), skupinami určenými na základě etnické příslušnosti (Soland, 2018) i mezi zeměmi. Ve výše zmíněné studii (Anaya & Zamarro, 2024) se v testu z matematiky podíl odpovědí poskytnutých s nedostatečným úsilím u 52 států pohyboval od 1 do 5,6 % (např. Finsko 1,8 %, Izrael 5,3 %) a mimo uvedený interval se nacházely pouze tři státy (Tunisko 6,5 %, Katar 9,2 % a Dominikánská republika 15,2 %). Uvedené poznatky vedou k úvaze, do jaké míry jsou srovnání různých skupin testovaných pouze na základě výsledného skóre v testu zkreslena nedostatečným úsilím, a to zejména v těch případech, kdy nelze vyloučit významné rozdíly v míře nedostatečného úsilí mezi srovnávanými skupinami.

V českém kontextu obdobné studie založené na analýze časových charakteristik chybí i přesto, že řada testů je zadávaných elektronicky (např. testování České školní inspekce, Scio, ale nově také testy v mezinárodních výzkumech TIMSS, PIRLS, PISA, ICILS) a log data by bylo možné získat a analyzovat. Řadu poznatků,

¹ Pro zjednodušení budeme v dalším textu používat pro označení zemí a ekonomik slovo státy.

44 například o rozdílech mezi dívkami a chlapci, lze získat ze zahraničních studií. O jiných aspektech, které plynou ze specifík českého systému školství, prozatím nemáme informaci.

Tato studie přispívá k poznání uvedené problematiky v souvislosti s jedním z potenciálně důležitých specifík, kterým je typ školy. Konkrétně se zabýváme nedostatečným úsilím při testování v základních školách a víceletých gymnáziích a zkoumáme, v jaké míře se objevuje v daných typech škol, jak se promítá do výsledků žáků a jaké může mít dopady při srovnávání žáků obou typů škol.

1 Metody identifikace a zohlednění rapid guessing behavior

Podstatou metod identifikace nedostatečného úsilí založených na časových charakteristikách je určení, resp. odhad výše diskutovaného minimálního času potřebného k odpovědi na položku. Ten je pak používán jako prahová hodnota (angl. *threshold*), na jejímž základě je klasifikována každá odpověď každého testovaného. Odpovědi poskytnuté za čas kratší, než je prahová hodnota, jsou klasifikovány jako projev rapid guessing behavior, ostatní pak jako projev solution behavior. Rozdíl mezi metodami spočívá ve způsobu, kterým prahovou hodnotu stanovují. Podle způsobu stanovení prahové hodnoty můžeme metody roztrdit do tří hlavních skupin:

1.1 Metody založené na externě stanovených kritériích

Jedná se o metody, které stanovují prahové hodnoty na základě vlastností položek, tedy nezávisle na získaných datech z daného testování.

Metoda pevné prahové hodnoty (*common threshold method*), někdy též nazývaná *common method* patří mezi nejjednodušší metody. Stanovuje prahovou hodnotu stejnou pro všechny položky či skupinu podobných položek a obvykle se jedná o čas v řádu jednotek sekund (více např. Kong et al., 2007), u kterého se předpokládá, že dostačuje pouze k aktu odpovědi (výběr odpovědi, příp. potvrzení).

Metoda času pro přečtení (*reading time method*) zohledňuje vlastnosti položek a při určování prahové hodnoty vychází z času, který je nutný k přečtení položky (Sahin & Colvin, 2020).

Metoda obsahu položky (*surface features method*) je obdobou metody času pro přečtení, vychází také z času nutného k přečtení položky, ale navíc zohledňuje i přítomnost netextových prvků jako jsou obrázky, grafy atp. (např. Wise, 2017).

Metody v této skupině však nezohledňují specifika položek z hlediska časové náročnosti na jejich pochopení či promyšlení odpovědi. Jako problematické se rovněž jeví i určení času nutného k přečtení položky, protože ten se mj. opírá o předpokládanou rychlost čtení, která může napříč různými podskupinami testovaných vykazovat značnou variabilitu (Lee & Haberman, 2015). U metody pevné prahové hodnoty je nejasný způsob určení konkrétní prahové hodnoty. Ve starších studiích

byl používán čas tři sekundy (např. Wise et al., 2004), v jiných byla prahová hodnota stanovena odhadem (*educated guess*), avšak není zcela jasné, na jakém základě byl odhad proveden (např. Huang et al., 2012).

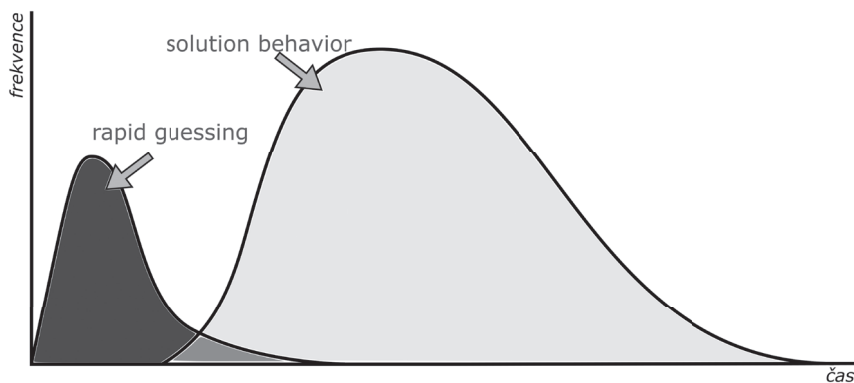
1.2 Metody založené na analýze časových údajů

Další metody již při stanovení prahové hodnoty vycházejí z časových údajů dané skupiny testovaných. Mohou tak zohlednit i náročnost položek na pochopení a promýšlení odpovědi, která se do časových údajů testovaných promítá.

Metoda normativní prahové hodnoty (*normative threshold method* nebo *normative method*), v literatuře uváděná též pod zkratkou *NT*, stanovuje prahové hodnoty pro každou položku zvlášť, a to na základě průměrného času odpovědi všech testovaných u dané položky. Prahová hodnota je stanovena jako určité procento průměrného času, obvykle jako 10, 15, příp. 20 %. Metoda předpokládá, že minimální čas a průměrný čas jsou v lineárním vztahu a je třeba stanovit jen příslušné procento (Wise & Ma, 2012).

Limitace metody normativní prahové hodnoty spočívá v její citlivosti na podíl rapid guessing behavior během testování, který se promítá do průměrného času odpovědi. Zejména u testování, kde je rapid guessing behavior častějším jevem, může být dopad na průměrný čas nezanedbatelný. Další nevýhodou je, že procento při stanovování prahových hodnot je fixní pro všechny položky a do výpočtu prahové hodnoty pak dále vstupuje již pouze průměr. Nejsou tak zohledněny rozdílné typy či vlastnosti položek, ani případná heterogenita skupiny testovaných, která se může specificky promítat ve tvaru rozdělení odpovědí v čase, protože průměr, jako míra centrální tendence, a fixní procento je nemohou zachytit.

Vizuální metoda (*visual method*), v literatuře označovaná též *visual inspection method* (případně jen zkratkou *VM*), je založena na vizuální analýze histogramu času u každé položky. Metoda předpokládá, že přítomnost rapid guessing behavior



Obrázek 1 Vizuální metoda – ilustrace bimodality v histogramu

46 způsobí bimodalitu histogramu s lokálním maximem u nejnižších časů (Kong et al. 2007). Předpokládaná bimodalita je způsobena tím, že odpovědi poskytnuté během rapid guessing behavior i solution behavior mají v čase v zásadě normální rozdělení, avšak jsou ve společném histogramu vzájemně posunuty (viz obrázek 1). Prahová hodnota pro danou položku je pak nastavena do lokálního minima vpravo od prvního vrcholu (Wise & DeMars, 2006).

Vizuální metoda není zatížena nevýhodami spojenými s určováním prahových hodnot odvozením z průměrného času jako u metody normativní prahové hodnoty a případný vyšší výskyt rapid guessing behavior se projevuje tím, že lokální maxima v histogramech jsou zřetelnější. Pokud je rapid guessing behavior naopak okrajovým jevem, nemusí být lokální maximum vůbec patrné a metodu nelze použít.

Metoda smíšeného logaritmicko-normálního rozdělení (*mixed lognormal distribution-based method*), v literatuře uváděná pod zkratkou *MLN* (podrobně Rios & Guo, 2020), předpokládá bimodální rozdělení odpovědí v čase, přičemž dolní vrchol představuje odpovědi bez dostatečného úsilí (rapid guessing behavior) a horní odpovědi s dostatečným vynaloženým úsilím (solution behavior). Metoda aplikuje automatizovaný proces, který využívá empirické rozdělení časů odpovědí, dosazuje smíšené logaritmicko-normální rozdělení a poté hledá nejnižší bod (průnik) mezi oběma mody rozdělení, který je následně stanoven jako prahová hodnota.

1.3 Metody založené na analýze časových údajů a podílu správných odpovědí

Jde o skupinu metod, které vycházejí z distribuce odpovědí v čase a přitom zohledňují podíl správných odpovědí. Základním východiskem je, že u odpovědí klasifikovaných jako projev rapid guessing behavior by měl podíl správných odpovědí odpovídat náhodné volbě.

Leeová a Jiaová (2014) uplatnily tento princip a navrhly metodu, kterou nazvaly *VITP*. Ta je založena na analýze úzkých časových úseků počínaje úsekem s nejkratšími časy. Prahová hodnota je pak nastavena na čas intervalu (levou stranu), u kterého podíl správných odpovědí vzroste nad úroveň odpovídající náhodné volbě a v dalších intervalech se nad touto úrovní drží.

Kumulativní proporcionální metoda (*cumulative proportional method*), v literatuře uváděná pod zkratkou *CUMP* (podrobně Guo et al., 2016), je modifikací metody *VITP*. Místo analýzy každého časového intervalu zvlášť bere metoda *CUMP* v úvahu všechny odpovědi od prvního intervalu až po aktuální zkoumaný interval. Pro každou kumulativní skupinu intervalů se pak porovnává přesnost (procento správných odpovědí) s pravděpodobností náhodného uhodnutí odpovědi. Práh je pak nastaven na čas intervalu (dolní mez), ve kterém přesnost odpovědí přesáhla míru odpovídající náhodné volbě.

Metoda detekce změny informace (*change in information method*), v literatuře uváděná pod zkratkou *ChInf* (podrobně Wise, 2019). Metoda pracuje s vteřinovými časovými úseky, které jsou vytvořeny zaokrouhlením doby odpovědi na nejbližší

sekundu. Pro odpovědi spadající do každého časového úseku se vypočítá korelace položky s testem (*item-total correlation*), tj. mezi správností odpovědi a celkovým skóre testu. Prahová hodnota je určena jako dolní hranice časového úseku, v němž korelace mezi položkou a celkovým skóre testu poprvé vykazuje konzistentní a trvalou hodnotu vyšší než 0,20 v průběhu více úseků (Wise, 2019).

Metoda detekce změny informace a přesnosti odpovědi (*change in information and accuracy method*), v literatuře uváděná pod zkratkou *ChIA* (podrobně Wise, 2019), je rozšířením předchozí metody o zohlednění přesnosti odpovědi (procenta správných odpovědi). Metoda určuje dvě pracovní prahové hodnoty. První dle metody *ChInf* a druhou jako dolní hranici časového úseku, v němž přesnost odpovědi začíná vykazovat trvalý nárůst směrem k hodnotě charakteristické pro *solution behavior*. Výsledná prahová hodnota je pak určena jako prostý průměr pracovních prahových hodnot.

Společným znakem metod založených na kombinaci času a podílu správných odpovědi je, že při zkoumání krátkých intervalů je detekce podílu správných odpovědi smysluplná pouze při dostatečném množství dat, tj. při dostatečném počtu testovaných. Současně musí být počet krátkých intervalů před potenciální prahovou hodnotou dostatečný k tomu, aby bylo možné zkoumat změny podílu správných odpovědi mezi intervaly. Problematickým aspektem uvedených metod je skutečnost, že „náhodná“ volba provedená člověkem neodpovídá z hlediska distribuce voleb náhodě z matematického pohledu, se kterou metody pracují. Řada studií ukázala, že při „náhodné“ volbě provedené člověkem jsou častěji vybírány středové hodnoty škály (Wise, 2017).

1.4 Zpracování odpovědi identifikovaných jako rapid guessing

Dalším krokem, který následuje po aplikaci výše shrnutých metod je rozhodnutí, jak zpracovat odpovědi, které jsou identifikovány jako *rapid guessing*. Literatura uvádí dva hlavní přístupy.

Filtrování dat. První přístup spočívá v tom, že analyzuje pouze data žáků, u kterých zastoupení odpovědi klasifikovaných jako *solution behavior* nekleslo pod určitou stanovenou mez. Je tedy použitelný pouze v případech, kdy není podstatné získat informaci o výkonu každého testovaného. Nachází tak využití například ve studiích zaměřených na srovnávání různých skupin testovaných na základě agregovaných výsledků a dosavadní výzkumy dokládají příznivé výsledky aplikace tohoto přístupu (např. Kong et al., 2007). K určení zmíněné meze je využíván konstrukt navržený Wisem a Kongovou (2005), nazvaný *response time effort (RTE)*, který vyjadřuje míru úsilí testovaného v celém testu:

$$RTE = \frac{N_{\text{solution behavior}}}{N_{\text{celkem}}},$$

kde $N_{\text{solution behavior}}$ je počet odpovědi respondenta klasifikovaných jako *solution behavior* a N_{celkem} je celkový počet položek v testu.

Analýza a srovnávání různých skupin jsou pak prováděny na datech testovaných, u kterých je $RTE \geq 0,9$ (např. Setzer et al., 2013, Soland, 2018), protože $RTE < 0,9$ indikuje celkově nízkou úroveň testového úsilí (Wise, 2015).

Korekce testového skóre. Druhou možností je vyřazení odpovědí, které byly klasifikovány jako rapid guessing a následné přepočítání testového skóre. Korekce testového skóre je založena na teorii odpovědi na položku (*item response theory – IRT*) a v literatuře je navržena řada různých přístupů (podrobněji např. Rios et al., 2017; Wise, 2017).

Korekce výsledků pomocí IRT je potřebná zejm. v situaci, kdy je cílem testování získat přesnější představu o znalostech každého testovaného. U studií zaměřených primárně na srovnávání různých podskupin na základě agregovaných výsledků je třeba tento přístup zvážit. Korekce výsledků pomocí IRT je totiž spojena s rizikem zkreslení, a to zejména u testovaných, u kterých rapid guessing nastal u většího počtu položek (Rios et al., 2017; Soland et al., 2019). Míra zkreslení pak závisí na tom, jak frekventovaným jevem rapid guessing behavior v dané studii je a jaké je procento testovaných, u nichž se toto chování projevilo ve větší míře.

2 Metody identifikace rapid guessing behavior v praxi

Problematika rapid guessing behavior je předmětem zájmu mnoha studií. Stále jsou vyvíjeny různé metody identifikace rapid guessing behavior a způsoby jeho zohlednění. V praxi je pak třeba při výběru konkrétních přístupů mít na zřeteli různá omezení a předpoklady, za nichž lze konkrétní metodu použít. U zvolené metody je vhodné následně věnovat pozornost jejím výstupům – prahovým hodnotám. Dosavadní zkušenosti totiž ukazují, že metody mohou u některých položek více či méně zjevně selhávat.

Soland et al. (2021) srovnával různé metody a zjistil, že metoda MLN určila u 13 % otázek extrémní prahovou hodnotu (> 100 sekund), přitom u ostatních typově stejných otázek byla prahová hodnota okolo 25 sekund. U metody CUMP nastal stejný jev u 3 % otázek. Zjistit, zda zvolená metoda selhala, však nemusí být vždy jednoduché, protože výsledná prahová hodnota nemusí být tak extrémní. Vše pak záleží na posouzení prahových hodnot v kontextu konkrétního testu a toho, zda a jak se liší prahové hodnoty u obdobných otázek (stejného typu, délky, srovnatelné obtížnosti apod.). Možné je také provést alespoň základní kontrolu histogramů v souladu s VM.

U metod pracujících s podílem správných odpovědí je vedle již zmíněné obtížnosti položek nutné zohlednit i počet testovaných. Wise a Maová (2012) upozorňují, že detekce nárůstu správných odpovědí (nad úroveň odpovídající náhodné volbě) je možná pouze při velkém množství dat, avšak bližší specifikaci neuvádějí. Konkrétnější představu lze získat až ze srovnávacích studií. Guová et al. (2016) aplikovali VM a CUMP na log data testu z matematiky na vzorku 1422 studentů a metody poskytly značně rozdílné prahové hodnoty, u VM to v průměru bylo 8,67 sekundy ($SD = 2,09$), zatímco v případě CUMP se jednalo o 14,96 sekundy ($SD = 8,90$). Naproti tomu Kröhne et al.

(2020) srovnávali metody (VM, VITP, CUMP, ChInf, ChIA) na větším vzorku (6751 studentů) a průměrné prahové hodnoty u testu z počítačové gramotnosti se pohybovaly mezi 8 a 11 sekundami. Wise (2019) ve své studii pracoval s ještě větším vzorkem (přes 23 000 studentů) a srovnávané metody (VM, VITP, upravená VITP, CUMP, ChInf, ChIA) poskytly průměrné prahové hodnoty v rozmezí 6,07 až 7,15 sekundy.

Mezi všemi diskutovanými metodami mají specifické postavení dvě z nich, a to VM a NT. VM vychází přímo ze základního předpokladu, že odpovědi poskytnuté během rapid guessing behavior i solution behavior mají v čase v zásadě normální rozdělení, avšak jsou ve společném histogramu vzájemně posunuty (viz obrázek 1). Zjišťování prahových hodnot je však u VM velmi pracné, časově náročné a možné jen tehdy, projeví-li se bimodalita. Pro využití v běžné vzdělávací praxi, zejm. u testů s mnoha položkami, tak není VM příliš praktická. I z toho důvodu navrhli Wise a Maová (2012) právě metodu NT (konkrétně v souvislosti s potřebou stanovit prahové hodnoty u adaptivních testů se stovkami položek). V průběhu let pak následovaly další metody, které zde byly diskutovány. Metoda NT je však stále využívána, přestože není tak přesná jako VM či jiné metody. Důvodem je nejen jednoduchost NT, ale především skutečnost, že jako jediná z metod založených na log datech nemá žádná omezení a prahovou hodnotu poskytne vždy. Nalezneme ji tak i u relativně aktuálních studií, které analyzují i rozsáhlé datové soubory, jako jsou data PISA (např. Anaya & Zamarro, 2024; Rios & Soland, 2022).

Metodologické výzvy představuje i samotná podstata identifikace nedbalého chování založená pouze na analýze časů, případně podílu správných odpovědí. To může vést k chybným klasifikacím – např. označení pomalé ale nedbalé odpovědi jako solution behavior a naopak rychlé odpovědi, avšak poskytnuté s adekvátním úsilím jako rapid guessing behavior (Wise, 2017). Aktuální přístupy se proto pokouší využít další log data (např. opakované čtení textu, návraty k položce, změna odpovědi) a začleňují je do analýz. Výsledky však ukazují, že přínos těchto indikátorů je zatím jen mírný a že jejich účinnost závisí na konkrétním testovém formátu i populaci (Welling et al., 2024).

3 Záměr studie a výzkumné otázky

Záměrem této studie je přispět k řešení uvedené problematiky v českém kontextu, a to prostřednictvím analýzy souvislostí rapid guessing behavior a výsledku v testu u podskupin žáků, definovaných na základě dvou znaků – typu školy a genderu.

Konkrétně si klademe následující výzkumné otázky:

- 1) Jaký je podíl rapid guessing behavior u různých skupin žáků (dle typu školy a genderu)?
- 2) Jaké jsou rozdíly ve výsledcích testu u jednotlivých skupin žáků před zohledněním rapid guessing behavior?
- 3) Jak se po zohlednění rapid guessing behavior změní rozdíly ve výsledcích v testu mezi skupinami?

4 Výzkumný vzorek a metody

Sběr dat probíhal v základních školách a víceletých gymnáziích v roce 2021. Výzkum se zaměřil na žáky 9. ročníků a odpovídajících ročníků víceletých gymnázií (ISCED 2). V rámci každého typu školy jsme provedli stratifikovaný výběr, přičemž stratifikační proměnnou byl region dle klasifikace NUTS 2 (*La Nomenclature des Unités Territoriales Statistiques*; Eurostat, 2021), tedy 8 oblastí, nikoli 14 krajů. V každém stratu byl uplatněn princip prostého náhodného výběru. Osloveno bylo postupně 152 škol a výzkumu se zúčastnilo 65 z nich (návratnost 42,8 %), z toho 35 základních škol (návratnost 41,5 %) a 30 víceletých gymnázií (návratnost 44,3 %). Školy byly instruovány, aby k účasti na výzkumu vyzvaly všechny žáky ze všech cílových ročníků (účast žáků byla dobrovolná).

Výzkumným nástrojem byla on-line webová aplikace, vyvinutá výzkumným týmem, která se skládala z dotazníkové části a testu z anglického jazyka. V dotazníkové části byly otázky týkající se demografických údajů, socioekonomického zázemí, postojů a zkušeností v souvislosti s výukou a učením se anglickému jazyku. Pro účely ověření znalostí anglického jazyka jsme použili rozřazovací test z anglického jazyka *Cambridge EFL Placement Test* (Cambridge University Press & Assessment, 2023), který obsahoval 25 otázek s výběrovou odpovědí (*multiple-choice item*), přičemž u prvních 15 otázek vybírali žáci ze tří nabízených odpovědí a u zbývajících 10 otázek ze čtyř nabízených odpovědí, z čehož byla vždy pouze jedna odpověď správná. Test neměl žádný časový limit. Webová aplikace byla navržena tak, aby vedle odpovědí na otázky zaznamenávala v dotazníkové i testové části řadu časových údajů, včetně časů jednotlivých voleb (výběr odpovědi, potvrzení odpovědi).

Za sběr dat ve škole byli odpovědní pouze pověřeni učitelé (byli za spolupráci honorováni) bez účasti výzkumníků. Před zahájením sběru dat dostali pověřeni učitelé podrobné pokyny a přesné znění instrukcí, které měli žákům přečíst. Každý žák mohl kdykoli přestat dotazník vyplňovat a své dosavadní odpovědi smazat. Žáci byli pověřenými učiteli informováni rovněž o tom, že součástí dotazníku je i test, který webová aplikace automaticky vyhodnotí, a každý tak na konci obdrží informaci o bodovém výsledku a své jazykové úrovni (dle *Common European Framework of Reference for Languages* – CEFR).

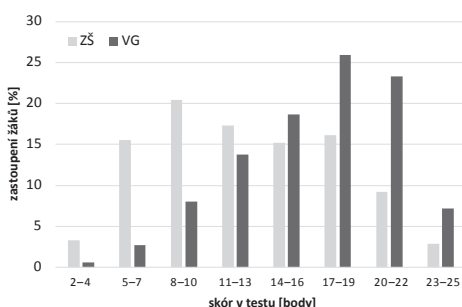
V tomto článku analyzujeme údaje žáků, kteří v rámci výzkumu dokončili test z anglického jazyka. Celkem se jednalo o 1344 žáků, z toho 631 ze základních škol a 713 z víceletých gymnázií.

5 Výsledky

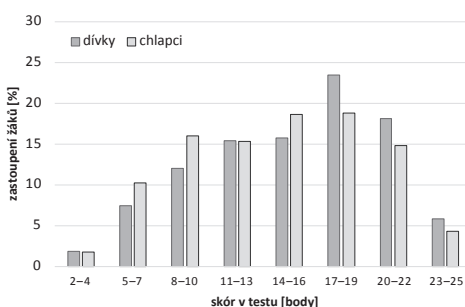
V této sekci nejprve seznamujeme s výsledky žáků v testu (před zohledněním rapid guessing behavior), pak zkoumáme časové charakteristiky aktivit žáků a jejich souvislosti s testovým skóre, následně se zabýváme podílem rapid guessing behavior u různých podskupin žáků a nakonec srovnáváme výsledky v testu před zohledněním rapid guessing behavior a po něm.

5.1 Výsledky žáků v testu

Základní výsledky žáků v testu z angličtiny jsou shrnuty v tabulce 1. Žáci dosáhli v testu průměrného výsledku 14,81 bodu (z 25 možných), přitom nejnižší dosažený výsledek byl 2 body a nejvyšší 25 bodů. Počty žáků dle distribuce výsledku v testu (obrázek 2) ukazují, že žáci víceletých gymnázií (dále jen VG) obsazují spíše vyšší bodové kategorie (s vrcholem v kategorii 17 až 19 bodů), zatímco žáci základních škol (dále jen ZŠ) naopak spíše nižší (s vrcholem 8 až 10 bodů). Podobně lze zobrazit i rozdíly testových skóre dle genderu (obrázek 3). Dívky převažují u vyšších bodových kategoriích a chlapci u nižších, avšak u obou skupin je nejvyšší četnost v kategorii 17 až 19 bodů.



Obrázek 2 Histogramy výsledku testu podle typu školy



Obrázek 3 Histogramy výsledku testu podle genderu

Z výsledků uvedených v tabulce 1 vyplývá, že žáci VG dosáhli vyššího průměrného bodového výsledku než žáci ZŠ (průměrný skóre VG je 16,61 a pro ZŠ je to 12,78). Velikost účinku měřená Cohenovým d indikuje střední efekt ($d = 0,79$). Rozdíl, i když menší a s malou věcnou významností ($d = 0,18$), je patrný i při srovnání dívek a chlapců (0,93 bodu).

Tabulka 1 Výsledky v testu pro skupiny žáků dle typu školy a genderu

	Výsledek v testu		Rozdíl průměrů	Cohen d
	průměr	sm. odch.		
Všichni žáci	14,81	5,18	–	–
Dle typu školy				
ZŠ	12,78	5,21	–3,83	–0,79
VG	16,61	4,45		
Dle genderu				
dívky	15,23	5,14	–0,93	0,18
chlapci	14,31	5,19		

Vzhledem k tomu, že data v naší studii mají vnořenou strukturu, resp. jsou hierarchicky uspořádána (škola – žák), aplikovali jsme pro zjištění statistické významnosti rozdílů mezi skupinami (dle genderu a typu školy) lineární smíšené modely s náhodnými efekty. Konkrétně jsme využili statistický software R a funkci *lme()* z balíčku *nlme* (Pinheiro et al., 2021). Do modelů jsme zahrnuli proměnnou kódující konkrétní školu.

Tabulka 2 Efekt genderu a typu školy

<i>Fixní efekty</i>	M1	M2	M3
Intercept	12,87*** (0,45)	14,89*** (0,41)	13,17*** (0,46)
Typ školy (VG)	3,63*** (0,64)		3,58*** (0,64)
Gender (chlapec)		-0,60* (0,24)	-0,58* (0,24)
<i>Náhodné efekty</i>			
SD škola	2,31	2,95	2,30
SD residual	4,34	4,32	4,33
AIC	7881	7900	7877
BIC	7901	7920	7903
logLik	-3936	-3946	-3933

Poznámka. * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$; $N = 1344$. SD = směrodatná odchylka; AIC = Akaikeovo informační kritérium; BIC = Bayesovské informační kritérium. Hodnoty v závorkách jsou standardní chyby odhadů.

Abychom prozkoumali, jaký vliv mají gender a typ školy na výsledné skóre v testu, odhadli jsme tři modely s náhodnými efekty (tabulka 2). První model (M1) obsahuje intercept, fixní proměnnou typ školy a náhodné efekty pro školu a žáka. Druhý model (M2) obsahuje místo typu školy gender. Třetí model (M3) obsahuje intercept, obě fixní proměnné (typ školy, gender) a náhodné efekty pro školu a žáka.

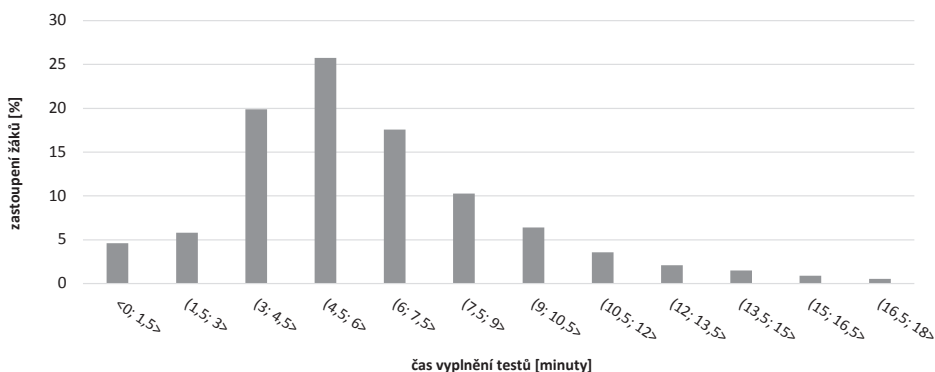
Z prvního modelu (M1) plyne signifikantní a poměrně silný efekt typu školy (3,63; $p < 0,001$), který je v případě gymnázií spojen s nárůstem průměrného skóre v testu. Naproti tomu efekt genderu (model M2) je výrazně menší, přesto je statisticky významný (-0,6; $p = 0,013$), a ukazuje, že u chlapců je spojen s nižším skóre v testu. Ve třetím modelu (M3) zahrnujícím oba prediktory, jsou jejich efekty mírně nižší, nicméně zůstávají signifikantní (efekt typu školy je 3,58; $p < 0,001$ a efekt genderu je -0,58; $p = 0,017$). Z modelů vyplývá, že i při zohlednění vzájemné závislosti pozorování v rámci škol jsou patrné statisticky významné rozdíly mezi skupinami dle genderu a typu školy.

5.2 Časové charakteristiky

Základní časové charakteristiky

Základní pohled na časové charakteristiky ilustruje histogram rozložení počtu žáků dle času vyplnění testu na obrázku 4. Průměrný čas vyplnění testu byl 6:18 (6 min. 18 s.), minimální byl 0:27 a maximální 31:56, přičemž většina žáků (95 %) dokončila celý test v čase kratším než 13 minut.

V histogramu je důležitý především první interval (0 až 1,5 min.), kde je zastoupeno 4,7 % žáků. Jedná se o skupinu, která celý test, všech 25 otázek, dokončila v čase do 1,5 min. (tj. v průměru 3,6 sekundy na otázku) a lze předpokládat, že v této skupině budou významně převažovat žáci, kteří nevěnovali testu téměř žádné úsilí a volili odpovědi většinou náhodně.



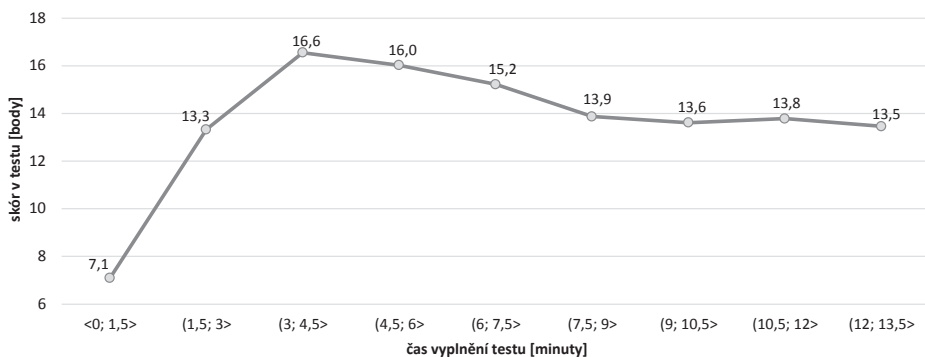
Obrázek 4 Histogram relativních četností žáků dle času vyplnění testu (interval 0 až 18 minut)

Pokud je uvedený předpoklad správný, pak by celkový výsledek této skupiny žáků měl být srovnatelný s hodnotou, kterou lze dosáhnout čistě náhodnou volbou. Pro test, který byl v našem výzkumu použit (25 otázek, z toho 15 otázek se třemi možnostmi odpovědi a 10 otázek se čtyřmi možnostmi odpovědi), za předpokladu stejných pravděpodobností zvolení nabízených odpovědí (0,25 u otázek se čtyřmi možnostmi, 0,33 se třemi), lze očekávat výsledné náhodné skóre 7,5 bodu. Vezme-li však v úvahu zjištění dosavadních studií, které ukázaly, že „náhodná“ volba provedená člověkem tenduje k preferenci „prostředních“ z nabízených odpovědí (Sahin & Colvin 2020; Wise, 2017 aj.), nemusí být výše uvedený předpoklad správný. Například Wise (2017) ve své studii zjistil, že u otázek se čtyřmi možnostmi odpovědi je při náhodné volbě možnost (a) volena testovanými ve 29 %, (b) 33 %, (c) 25 % a možnost (d) pouze ve 13 % případů (výsledky se však pro různé typy otázek lišily).

Z toho důvodu byla provedena analýza voleb zvlášť pro každou otázku, a to na základě odpovědí, které byly poskytnuty do 1,5 sekundy, tedy v tak krátkém čase,

54 že by nebylo možné přechíst žádnou z 25 otázek testu ani rychločtením (při rychlosti 500 slov/min.). Zjištěné rozložení voleb u otázek se třemi možnostmi odpovědi bylo (a) 38 %, (b) 42 %, (c) 19 % a u otázek se čtyřmi možnostmi odpovědi bylo (a) 30 %, (b) 32 %, (c) 27 % a (d) 11 %. Očekávané skóre, které lze dosáhnout volbou provedenou žákem zcela bez zvažování otázky i nabízených odpovědí, je pro námi použitý test 6,7 bodu (dále jen vypočtené *náhodné skóre*).

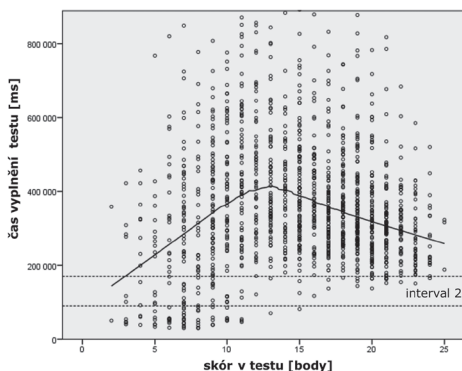
Výsledky žáků v testu pro všechny časové intervaly ukazuje následující graf (obrázek 5). Skupina žáků spadajících do prvního časového intervalu (0 až 1,5 min.) v průměru dosáhla 7,1 bodu. Vezmeme-li v úvahu průměrný výsledek v testu (14,81 bodu) a výsledky žáků ve všech ostatních intervalech (obrázek 5), můžeme konstatovat, že výsledek prvního intervalu je ojedinělý, ve srovnání s ostatními intervaly i celkovým průměrem a srovnatelný s vypočteným *náhodným skóre* (6,7 bodu). Pro další analýzy lze tedy vycházet z předpokladu, že mezi žáky z prvního intervalu významně převažují ti, kteří nevěnovali testu téměř žádné úsilí a volili odpovědi většinou náhodně.



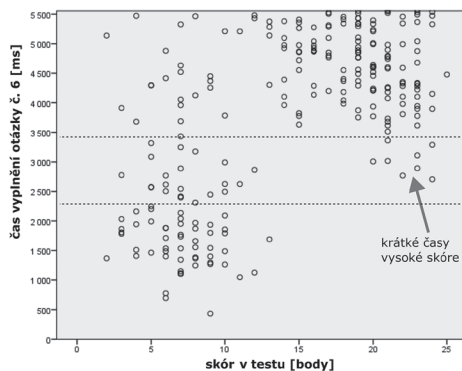
Obrázek 5 Výsledky v testu podle času jeho vyplnění

Čas a výsledek v testu. Pokud jde o vztah času vyplnění testu a výsledku v testu, ukazuje křivka v grafu na obrázku 6, že od nejnižšího bodového skóre v testu směrem ke 13 bodům roste čas vyplnění testu a následně, směrem k nejvyššímu bodovému skóre naopak čas zase postupně klesá.

Uvedené zjištění má s ohledem na zamýšlenou aplikaci metod identifikace rapid guessing behavior význam především v souvislosti s druhým časovým intervalem (1,5 až 3 min.). V bodovém grafu (obrázek 6) je druhý interval vyznačen (interval 2) a větší část žáků, která do něj spadá, je vlevo. Lze předpokládat, že v levé skupině, vzhledem ke krátkým časům vyplnění testu a skóre blízkému vypočtenému *náhodnému skóre*, bude zastoupení rapid guessing behavior stále nadprůměrné. Nicméně další část žáků z druhého intervalu (1,5 až 3 min.) se nachází vpravo u vysokých bodových výsledků (21 až 25 bodů). Nízké časy by napovídaly přítomnost rapid guessing behavior, ale poměrně vysoké skóre svědčí o opaku, neboť u žáků s tak vysokým skóre



Obrázek 6 Vztah mezi výsledkem v testu a časem vyplnění testu s křivkou konstruovanou metodou



Obrázek 7 Vztah mezi výsledkem v testu a časem vyplnění otázky 6 (detail do 5,5 s) s vyznačenou skupinou žáků s krátkým časem a vysokým skóre

v testu je pravděpodobnost náhodného odpovídání zanedbatelná. Jedná se tedy o skupinu žáků, která je schopna správně, a přitom velmi rychle, odpovídat, což lze interpretovat jako doklad jejich nadprůměrných znalostí angličtiny.

Stejná analýza byla provedena až na úroveň jednotlivých otázek (viz obrázek 7, ukázka otázky 6). Skupina žáků s nízkými časy a vysokým skóre je přítomna u všech otázek a u všech kromě první je oddělena od skupiny žáků se skóre kolem vypočteného *náhodného* skóre. Popsaný jev lze využít při stanovení prahových hodnot. Na jeho základě je na úrovni otázek možné nastavit orientační kontrolní hranici, kterou by prahová hodnota neměla překročit, aby nezasáhla do skupiny žáků, u nichž vysoké skóre ukazuje na převládající solution behavior.

5.3 Identifikace a zohlednění rapid guessing behavior

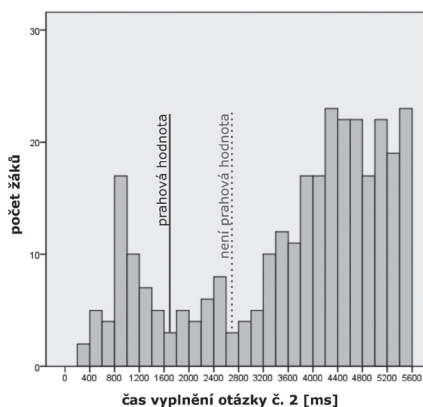
Na základě analýz časových charakteristik, jejich souvislostí s výsledkem v testu a souvislostí se skupinami dle genderu a typu školy a se zřetelem k vlastnostem metod identifikace rapid guessing behavior (viz část 1) a po zjištění, že se v distribuci odpovědí objevuje bimodalita (podrobněji níže), byla pro identifikaci rapid guessing behavior využita vizuální metoda.

Metody založené na externě stanovených kritériích nebyly využity, protože nezohledňují specifika položek z hlediska časové náročnosti na jejich pochopení či promyšlení odpovědi ani specifika skupiny testovaných. To může být problematické zejména u otázek, které jsou poměrně krátké a u nichž je čas nutný k promyšlení odpovědi srovnatelný či dokonce delší než čas nutný k prostému přečtení otázky a nabízených odpovědí. Metoda normativní prahové hodnoty nebyla využita, protože výše zmíněná analýza nasvědčuje, že v námi zkoumaném vzorku je z hlediska rapid guessing behavior mezi skupinami poměrně značná heterogenita. Metody kombinující čas

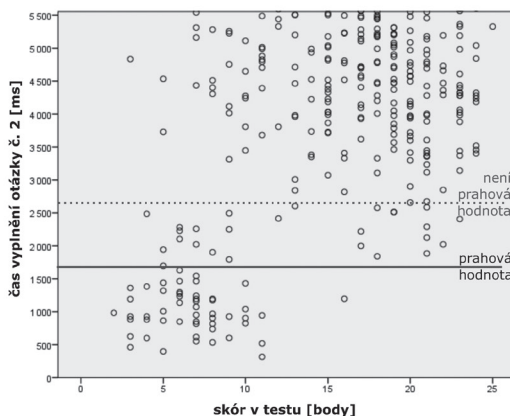
a podíl správných odpovědí využity být nemohly z více důvodů. Hlavním důvodem je jejich omezení, které spočívá v tom, že nelze definovat prahové hodnoty pro obtížné položky, u nichž je celkový podíl správných odpovědí na úrovni odpovídající náhodné volbě (Rios & Guo, 2020; Wise, 2019). Některé z metod (VITP, CUMP, ChInf, ChIA) nemají dobré výsledky nejen pro obtížné položky, ale i pro příliš snadné a v takových případech vrací velmi vysoké – nebo naopak velmi nízké – prahové hodnoty (Guo et al., 2016; Wise, 2019). Rozřazovací test, který v naší studii používáme, má však ze své podstaty jak položky velmi snadné, tak i položky velmi obtížné. Zvažována byla i aplikace metody smíšeného logaritmicko-normálního rozdělení (MLN), která pracuje pouze s distribucí časů odpovědí a podíl správných odpovědí nezohledňuje, avšak ta se v praxi ukázala jako problematická (viz oddíl 2).

Vizuální metoda stanovuje prahové hodnoty pro každou otázku na základě vizuální inspekce histogramů distribuce odpovědí v čase a zjišťuje, zda je rozdělení bimodální s prvním vrcholem u nízkých časů. Aby vizuální metoda přinesla žádané výsledky musí být rapid guessing behavior zastoupeno v takové míře, aby se bimodalita mohla projevit. Tato podmínka je v našem výzkumu splněna, neboť jak bylo ukázáno v předchozí části, 4,7 % žáků dokončilo celý test v čase kratším než 1,5 minuty, jejich průměrný výsledek je srovnatelný s vypočteným *náhodným* skóre a předpokládáme, že v této skupině rapid guessing významně převládá. Další důležitou otázkou je určení časových intervalů histogramu tak, aby byla případná bimodalita zřetelná. Příliš dlouhé intervaly distribuci vyhladí a příliš krátké se projeví mnoha lokálními maximy.

V našem výzkumu jsme postupně zkoumali intervaly v rozsahu 50 až 400 ms a bimodalita se objevila u 24 z 25 otázek. Neobjevila se pouze u první otázky, což může být dáno tím, že žáci se v čase, který aplikace naměřila, neseznamovali jen s otázkou, ale vlastně poprvé i s podobou testu. U ostatních otázek se jako vhodné ukázaly intervaly v rozmezí 150 až 250 ms. Prahové hodnoty pak byly nastaveny



Obrázek 8 Histogram času vyplnění druhé otázky testu (detail do 5,6 s)

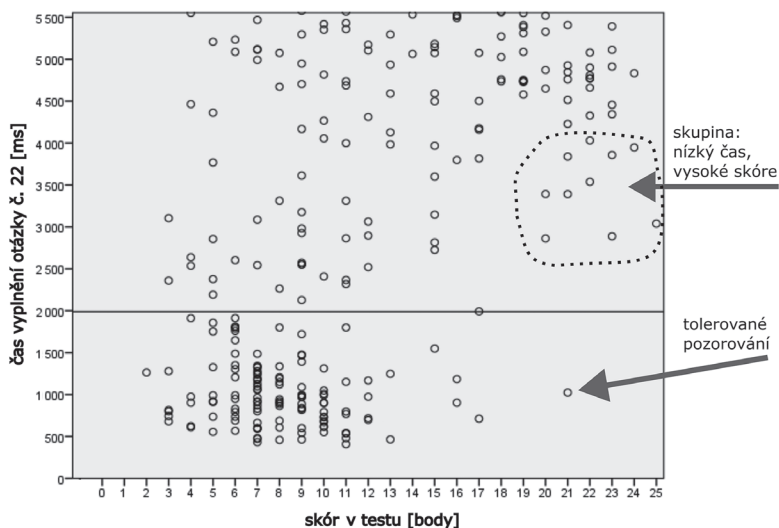


Obrázek 9 Bodový graf skóre v testu a času vyplnění druhé otázky testu (detail do 5,5 s)

do středu intervalů, které odpovídaly lokálnímu minimu za vrcholem distribuce u nízkých časů.

U některých otázek se však objevily u nízkých časů dva vrcholy (viz obrázek 8, ukázka otázky 2). Proto byla u těchto otázek provedena vizuální analýza bodových grafů zobrazujících čas odpovědi na otázku a výsledek v testu (obrázek 9, ukázka otázky 2). U všech otázek s dvěma vrcholy u nízkých časů se ukázalo, že ve druhém vrcholu se objevují žáci s výsledným skóre 21 až 25 bodů. Vzhledem k tomu, že u těchto žáků se rapid guessing předpokládá pouze v minimální míře, je druhý vrchol zřejmě místem, kde dochází k překryvu distribuce žáků s rapid guessing behavior a žáků se solution behavior (překryv viz obr. 1). Pro minimalizaci toho, že jako rapid guessing budou označeny i odpovědi, které jsou ve skutečnosti produktem solution behavior, byly prahové hodnoty nastaveny již za prvním vrcholem.

Následně bylo i u dalších otázek (s jedním vrcholem distribuce u nízkých časů) kontrolováno, zda prahová hodnota nezasahuje do skupiny žáků, jejichž výsledek v testu se pohyboval v rozmezí 21 až 25 bodů. Vycházelo se ze stejného předpokladu, že skupina žáků s vyššími bodovými výsledky se k rapid guessing behavior uchyluje pouze v minimální míře a jejich rychlá odpověď je spíše projevem nadprůměrných znalostí. Kontrola u všech otázek ukázala, že prahové hodnoty nezasahují do dané skupiny. Objevily se pouze izolované případy odpovědi s časem nižším než prahová hodnota (viz obrázek 10, ukázka otázky 22), a to pouze u sedmi otázek (u 6 z nich vždy 1 případ a u 1 otázky 2 případy). Tyto případy však byly tolerovány, protože ani u skupiny žáků s vyššími bodovými výsledky nelze rapid guessing u některé otázky zcela vyloučit. Práhové hodnoty se pro jednotlivé otázky pohybovaly v rozmezí 1,4 až 2,8 sekundy, v průměru to bylo 2,1 sekundy.

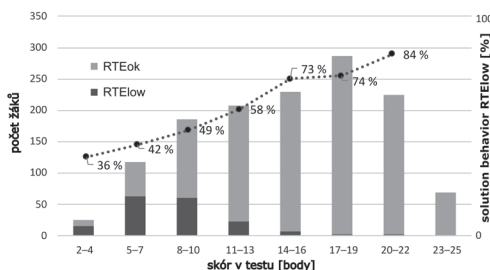


Obrázek 10 Bodový graf skóre v testu a času vyplnění otázky 22 (detail do 5,5 s)

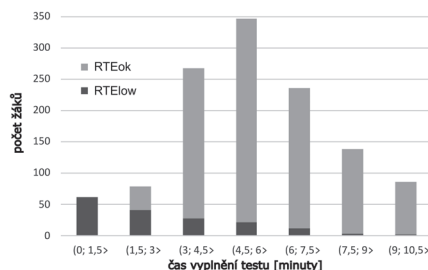
Dalším krokem aplikace vizuální metody byla klasifikace všech odpovědí všech žáků. Odpovědi poskytnuté za čas kratší než prahová hodnota byly označeny jako projev rapid guessing behavior a odpovědi ostatní jako produkt solution behavior s tím, že v případě první otázky, kde se bimodalita neprojevila, byly všechny odpovědi klasifikovány jako solution behavior. Pro každého žáka byl pak spočítán RTE a do srovnávacích analýz podskupin žáků (dle typu školy a genderu) byli na základě doporučení (např. Setzer et al., 2013; Soland, 2018) zahrnuti respondenti, u kterých bylo $RTE \geq 0,9$ (podíl solution behavior dosáhl alespoň 90 %).

Z celkového počtu 1344 žáků splnilo podmínku ($RTE \geq 0,9$) celkem 1173, tj. 87,3 %. Tuto skupinu žáků budeme dále označovat zkratkou RTEok. Vzhledem k celkovému počtu otázek testu (25) a nastavené hranici ($RTE \geq 0,9$) se žáci RTEok mohli uchýlit k rapid guessing behavior nejvýše u dvou otázek. Zbývajících 171 žáků (12,7 %) stanovenou podmínku nesplnilo ($RTE < 0,9$) – uchýlili se k rapid guessing behavior nejméně u tří otázek. Dále budeme tuto skupinu označovat jako žáky s podlimitním RTE, resp. zkratkou RTElow.

Skupina žáků s podlimitním RTE. Žáci s podlimitním RTE (RTElow) dosáhli v testu v průměru pouze 8,33 bodu. Jejich distribuce dle bodového výsledku v testu (obrázek 11) ukazuje, že největší dopad má aplikace vizuální metody v kategorii 2 až 4 body (60 % jsou žáci RTElow) a směrem k vyšším bodovým kategoriím jejich podíl postupně klesá. Graf současně ukazuje i průměrný podíl solution behavior u RTElow v jednotlivých bodových kategoriích, který směrem vpravo narůstá. Detailní analýza tohoto jevu, konkrétně zastoupení solution behavior u jednotlivých otázek zvláště pro každou bodovou kategorii, ukázala na změny chování RTElow v průběhu vyplňování testu. Zatímco u RTElow z první kategorie (2 až 4 body) je solution behavior podobně nízké u všech otázek (nedostatečné úsilí už od prvních otázek a výrazně se nemění), u dalších kategorií lze sledovat, že úsilí žáků na začátku testu bylo vyšší a od určité otázky zřetelně pokleslo. Například v kategorii 11 až 13 bodů bylo solution behavior u prvních 11 otázek v průměru 89 % (86 % až 95 %) a od otázky 12 došlo k výraznému poklesu průměru solution behavior na 34 % (s jednou výjimkou bylo solution behavior u všech 14 % až 50 %). Z toho lze usuzovat, že jistá část žáků RTElow, přestala



Obrázek 11 Histogram výsledku testu pro skupiny RTEok a RTElow a míra solution behavior skupiny RTElow



Obrázek 12 Histogram času vyplnění testu pro skupiny RTEok a RTElow

v určité části testu vynakládat dostatečné úsilí a u zbývajících otázek se častěji uchýlila k rapid guessing behavior.

U žáků spadajících do RTElow je také zajímavá souvislost s výše diskutovanými časovými charakteristikami (viz část Základní časové charakteristiky), především souvislost se skupinou žáků, kteří spadají do prvního časového intervalu (dokončili celý test do 1,5 minuty). Ukázalo se, že všechna pozorování z prvního intervalu se týkají žáků skupiny RTElow a žádného ze skupiny RTEok (viz obrázek 12).

Struktura skupiny žáků s podlimitním RTE. Procento žáků skupiny RTElow a jejich zastoupení ve vzorku dle genderu a typu školy shrnuje tabulka 3. Do skupiny RTElow spadá téměř čtvrtina žáků ZŠ (23,6 %), ale pouze minimum žáků VG (3,1 %). Z hlediska genderu je mezi žáky RTElow dvakrát více chlapců (17,5 %) než dívek (8,8 %). Uplatníme-li obě hlediska současně, pak největší podíl žáků s podlimitním RTE (29,9 %) je v podskupině chlapců ze ZŠ a naopak nejméně u dívek z VG (1,7 %).

Tabulka 3 Zastoupení žáků s podlimitním RTE dle typu školy a genderu (v %)

Typ školy	Dívky	Chlapci	Celkem
ZŠ	17,7	29,9	23,6
VG	1,7	5,0	3,1
Celkem	8,8	17,5	12,7

Poznámka. Hodnota 29,9 % znamená procento žáků ze skupiny chlapců ZŠ, kteří patří do skupiny RTElow.

Výsledky v testu po zohlednění rapid guessing behavior. V tabulce 4 jsou uvedeny výsledky v testu po zohlednění rapid guessing behavior (žáci RTEok) a současně pro srovnání zrekapitulovány i výsledky původní (všichni žáci). Průměrný výsledek v testu žáků RTEok dosáhl hodnoty 15,76 bodu a ve srovnání s výsledkem všech žáků (14,81) je o 0,95 bodu vyšší. U chlapců je výsledek vyšší o 1,22 bodu (ze 14,31 na 15,53) a u dívek o 0,7 bodu (z 15,23 na 15,93). Z hlediska typu školy je skóre u ZŠ vyšší o 1,42 bodu (z 12,78 na 14,20) a v případě VG pouze o 0,24 bodu (ze 16,61 na 16,85).

Tabulka 4 Výsledky všech žáků a výsledky RTEok v testu pro podskupiny dle typu školy a genderu

	Všichni žáci				Žáci RTEok			
	průměr	sm. odch.	rozdíl	Cohen <i>d</i>	průměr	sm. odch.	rozdíl	Cohen <i>d</i>
Všichni žáci	14,81	5,18	–	–	15,76	4,72	–	–
Dle typu školy								
ZŠ	12,78	5,21	–3,83	–0,79	14,20	4,91	–2,65	–0,59
VG	16,61	4,45			16,85	4,25		
Dle genderu								
dívky	15,23	5,14	0,92	0,18	15,93	4,73	0,40	0,09
chlapci	14,31	5,19			15,53	4,69		

Výsledky žáků RTEok v jednotlivých podskupinách mají vedle informace o úrovni znalostí anglického jazyka podstatné implikace i při jejich srovnávání. Jak je patrné z tabulky 4, jsou rozdíly mezi podskupinami dle genderu a typu školy u RTEok nižší, než byly u všech žáků. Stále platí, že průměrný bodový výsledek žáků VG (16,85) je vyšší než u žáků ZŠ (14,20), ale rozdíl je nižší (2,65) a věcná významnost měřená Cohenovým d , přestože jde stále o střední efekt, poklesla zhruba o čtvrtinu ($d = 0,59$). Při srovnání průměrného skóre dívek (15,93) a chlapců (15,53) je rozdíl již pouze 0,4 bodu, Cohenovo d kleslo na polovinu ($d = 0,09$), a rozdíl je tak věcně nevýznamný.

Pro zjištění statistické významnosti rozdílů mezi skupinami (dle genderu a typu školy) u RTEok jsme opět aplikovali lineární smíšené modely s náhodnými efekty. V tabulce 5 jsou modely, vypočtené na základě dat pro žáky RTEok, označeny MOK1, MOK2 a MOK3. Současně pro jsou pro srovnání v tabulce uvedeny i původní modely (M1, M2, M3) počítané na základě všech pozorování.

Tabulka 5 Efekt genderu a typu školy v modelech pro všechny žáky a modelech pro žáky z RTEok

	Modely pro všechny žáky ($N = 1344$)			Modely pro RTEok ($N = 1173$)		
<i>Fixed effects</i>	M1	M2	M3	MOK1	MOK2	MOK3
Intercept	12,87*** (0,45)	14,89*** (0,41)	13,17*** (0,46)	14,17*** (0,42)	15,61*** (0,36)	14,33*** (0,43)
Typ školy (VG)	3,63*** (0,64)		3,58*** (0,64)	2,58*** (0,59)		2,56*** (0,59)
Gender (chlapec)		-0,60* (0,24)	-0,58* (0,24)		-0,36 (0,25)	-0,33 (0,25)
<i>Random effects</i>						
SD škola	2,31	2,95	2,30	2,03	2,42	2,04
SD residual	4,34	4,32	4,33	4,12	4,12	4,12
AIC	7881	7900	7877	6760	6775	6760
BIC	7901	7920	7903	6780	6795	6785
logLik	-3936	-3946	-3933	-3376	-3383	-3375

Poznámka. * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$. SD = směrodatná odchylka; AIC = Akaikeovo informační kritérium; BIC = Bayesovské informační kritérium. Hodnoty v závorkách jsou standardní chyby odhadů.

Srovnáme-li původní a nové hodnoty efektů, je zřejmé, že efekt typu školy oslabil, avšak je stále poměrně silný a signifikantní (původní efekt je 3,63; $p < 0,001$ a efekt pro RTEok je 2,58; $p < 0,001$). Lze tedy konstatovat, že typ školy VG je statisticky významným prediktorem vyššího skóre v testu. Naproti tomu efekt genderu již v modelech pro RTEok není statisticky významný (původní efekt je -0,6; $p = 0,013$

a efekt pro RTEok je $-0,36$; $p = 0,150$). Z toho lze usoudit, že statisticky významné rozdíly dané genderem v modelech pro všechny žáky byly důsledkem rapid guessing behavior, resp. jeho vyšší incidencí u chlapců (viz tabulka 3).

5.4 Shrnutí

Naše studie ukázala, že míra rapid guessing behavior při testování znalostí z anglického jazyka souvisí s genderem žáků a s typem školy, kterou navštěvují.

Rapid guessing behavior u tří a více otázek v testu se u žáků základních škol v naší studii objevuje téměř osmkrát častěji než u žáků víceletých gymnázií. Mezi těmi, kteří projevili rapid guessing behavior u více než poloviny otázek testu, je zastoupení žáků základních škol dokonce třináctkrát větší než u žáků víceletých gymnázií.

Analýza z hlediska genderu žáků rovněž poukázala na významné rozdíly. Chlapců, kteří se uchýlili k rapid guessing behavior u tří a více otázek, je dvakrát více než dívek. Na rozdíl od srovnání dle typu školy je tento poměr obdobný i u žáků, kteří projevili rapid guessing behavior u většiny otázek testu (poměr chlapců a dívek dosáhl hodnoty 1,8).

Zohlednění rapid guessing behavior se promítlo i do srovnávání výsledků v testu dle genderu a typu školy. Jestliže původní výsledky ukázaly statisticky významné rozdíly dle genderu i typu školy, po zohlednění rapid guessing behavior zůstaly signifikantní jen rozdíly dané typem školy. Nelze tedy vyloučit, že skutečné kompetence chlapců a dívek jsou v angličtině srovnatelné a rozdíly původních výsledků testování (všech žáků) jsou v low-stakes podmínkách způsobeny převážně vyšší mírou rapid guessing behavior u chlapců.

6 Diskuze

Poznatky z výzkumů zabývajících se srovnáváním míry nedbalého odpovídání, resp. rapid guessing behavior na různých typech českých škol, a navíc v souvislosti s měřením znalostí anglického jazyka, zatím chybí. Možné vysvětlení rozdílů zjištěných v této studii lze tedy provést pouze z hledisek, která vycházejí z charakteristik obou typů škol, zejména pak z mechanismu, kterým jsou žáci přijímáni na víceletá gymnázia. U žáků, kteří byli přijati na víceletá gymnázia, můžeme s ohledem na parametry a náročnost přijímacího procesu předpokládat větší smysl pro plnění studijních povinností i vyšší schopnost vynaložit úsilí při aktivitách souvisejících se studiem. Svou roli mohou sehrávat i dosavadní zkušenosti žáků s testováním. Dále je třeba brát v úvahu i prostředí školy. To je jistě ovlivněno samotnými žáky, avšak způsob výběru homogenizuje skupinu žáků právě z hlediska charakteristik, které jsou relevantní v přijímacím řízení (VG). Tyto charakteristiky, přístup k povinnostem a schopnost vynaložit úsilí mohou vysvětlovat výrazně menší podíl rapid guessing behavior na víceletých gymnáziích.

Pokud jde o odlišnosti v míře rapid guessing behavior u dívek a chlapců zjištěné v naší studii, lze vysvětlení čerpat ze zahraniční literatury. Studie, které se zabývají low-stakes testy, ukazují, že dívky jsou přívětivější a svědomitější (*agreeable or conscientious*), přičemž tyto charakteristiky pak spojují s větší ochotou zodpovědně spolupracovat při testování (např. DeMars et al., 2013). Předmětem diskusí je i celkový přístup dívek a chlapců ke studijním aktivitám a motivaci skládat testy (více např. Kröhne et al., 2020; Zhao et al., 2022). Motivaci složit test lze považovat za podstatný prediktor vynaložení dostatečného úsilí a řada studií přináší zjištění, že dívky jsou v tomto směru motivovány více. Například u švédských 14–15letých žáků, kteří se účastnili šetření TIMSS, byla motivace k psaní testů přibližně o 0,33 směrodatné odchylky vyšší u dívek než u chlapců (Eklöf, 2007).

Motivace odpovídat v testu je dále ovlivněna psychologickými faktory, které popisuje například teorie očekávání a hodnoty (*expectancy-value theory*). Ta předpokládá, že míra úsilí, které jedinec vynakládá, závisí na tom, jaký význam činnosti přikládá a jaké má očekávání ohledně jejího výsledku (Wigfield & Eccles, 2000). Pokud žák považuje test za nezajímavý, zbytečný nebo příliš zatěžující, jeho ochota se soustředit a odpovídat odpovědně se snižuje – bez ohledu na jeho faktické schopnosti (Penk & Schipolowski, 2015). Pokles úsilí během testu může být způsoben i únavou, ztrátou pozornosti nebo nudou. Opakovaně bylo prokázáno, že respondenti častěji odpovídají nedbale u položek nacházejících se ke konci testu (Goldhammer et al., 2017; Pools & Monseur, 2021). Důležitou roli sehrávají také charakteristiky samotných položek. Položky, které působí vizuálně zahlcujícím dojmem (např. příliš dlouhé zadání, složité formátování), bývají spojeny s nižší mírou snahy ze strany žáků (Wise et al., 2009). Také umístění položky v testu, délka textu či monotónní formát po sobě jdoucích položek mohou vést k tomu, že žáci neodpovídají s dostatečným úsilím (Goldhammer et al., 2017; Wolf et al., 1995).

Celkově lze říci, že nedbalé odpovídání má multifaktoriální charakter a jeho výskyt je ovlivněn vlastnostmi žáků, charakteristikami testu, způsobem jeho administrace a mnoha dalšími činiteli. Identifikace a zohlednění nedbalých odpovědí jsou proto klíčovými kroky pro zajištění validní interpretace výsledků žákovského výkonu v testových šetřeních.

7 Limity studie

Hlavní limity studie spočívají především v použité vizuální metodě identifikace rapid guessing behavior. Výsledná prahová hodnota je ovlivněna zvolenou délkou časových intervalů histogramu, který je pak předmětem vizuální inspekce realizované výzkumníkem (Lee & Jia, 2014; Rios et al., 2017). S touto limitací jsme se snažili vypořádat zkoumáním a porovnáváním histogramů s různými délkami časových intervalů s cílem nalézt takovou délku, která vedla k nejzřetelnějšímu vizuálnímu projevu bimodality. Následně jsme provedli i kontrolu, zda prahové hodnoty nezasahují do

skupiny žáků s velmi nadprůměrnými výsledky v testu, u kterých je rapid guessing nepravděpodobný.

Další limitace jsou spojeny se způsobem zohlednění odpovědí označených vizuální metodou jako produkt rapid guessing behavior. Způsob zvolený v naší studii (analýza dat žáků s RTE $\geq 0,9$) stojí na předpokladu, že rapid guessing behavior nesouvisí s reálnými kompetencemi testovaného. V literatuře se však v tomto směru objevují odlišná zjištění. Rios et al. (2017) upozorňují, že u testů, kde tento předpoklad není splněn, může dojít ke zkreslení agregovaných výsledků. Příkladem mohou být testy, kde postupně graduje obtížnost položek až na úroveň, kterou testovaný s malými znalostmi vyhodnotí jako pro něj velmi náročnou a další snahu vynakládat potřebné úsilí vyhodnotí jako neefektivní, rezignuje a uchýlí se k rapid guessing behavior. Jeho chování pak lze spíše interpretovat jako důsledek a doklad nižší úrovně znalostí než ochoty vynaložit adekvátní úsilí během testování. Naopak Wise a Maová (2012) uvádějí, že rapid guessing relativně nesouvisí se schopnostmi testovaného. Vyšší incidence nedbalého odpovídání u obtížných položek může být dána jeho nižší ochotou. Wise (2017) popisuje, že pokud testovaný vyhodnotí položku jako příliš náročnou ve vztahu ke své aktuální ochotě vynaložit úsilí, rozhodne se pro nedbalé odpovídání (rapid guessing). Toto rozhodnutí testovaného je výsledkem porovnání mezi požadavky položky (*resource demands*) a aktuální disponibilní ochotou vynaložit úsilí (*effort capacity*; Wise, 2017).

Poděkování

Studie je výsledkem výzkumné činnosti podporované Grantovou agenturou České republiky v rámci grantu GA ČR 24-10968S „Vývoj motivace k učení se anglickému jazyku u českých žáků z různých typů škol na úrovni vyššího sekundárního vzdělávání“.

Literatura

- Anaya, L. M., & Zamorro, G. (2024). The role of student effort on performance in PISA: revisiting the gender gap in achievement. *Oxford Economic Papers*, 76(2), 533–560. <https://doi.org/10.1093/oep/gpad018>
- Cambridge University Press & Assessment. (2023, 20. dubna). *Cambridge English: Test your English – for schools*. <https://www.cambridgeenglish.org/test-your-english/for-schools/>
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, 8, 69–82.
- Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108(19), 7716–7720. <https://doi.org/10.1073/pnas.1018601108>
- Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, 7(3), 311–326. <https://doi.org/10.1080/15305050701438074>

- Eurostat. (2021, 5. prosince). *NUTS – Nomenclature of territorial units for statistics*. <https://ec.europa.eu/eurostat/web/nuts/background>
- Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, 5(1), Article 18. <https://doi.org/10.1186/s40536-017-0051-9>
- Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, 29(3), 173–183. <https://doi.org/10.1080/08957347.2016.1171766>
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Jacoby, W. G. (2000). Loess: A nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4), 577–613. [https://doi.org/10.1016/s0261-3794\(99\)00028-1](https://doi.org/10.1016/s0261-3794(99)00028-1)
- Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, 67(4), 606–619. <https://doi.org/10.1177/0013164406294779>
- Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014, duben). *A strategy for increasing student motivation on low-stakes assessments* [Paper presentation]. Annual Meeting of the American Educational Research Association, Philadelphia, USA.
- Kröhne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, 62(2), 147–177. <https://doi.org/10.25656/01:23630>
- Lee, Y.-H., & Haberman, S. J. (2015). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, 16(3), 240–267. <https://doi.org/10.1080/15305058.2015.1085385>
- Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, 2, Article 8. <https://doi.org/10.1186/s40536-014-0008-1>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437–455. <https://doi.org/10.1037/a0028085>
- Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences*, 42, 27–35. <https://doi.org/10.1016/j.lindif.2015.08.002>
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). *nlme: Linear and non-linear mixed effects models*. <https://CRAN.R-project.org/package=nlme>
- Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, 9(1), Article 10. <https://doi.org/10.1186/s40536-021-00104-6>
- Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, 27(4), 356–373. <https://doi.org/10.1080/10627197.2022.2110465>
- Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, 33(4), 263–279. <https://doi.org/10.1080/08957347.2020.1789141>
- Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: to filter unmotivated examinees or not?. *International Journal of Testing*, 17, 74–104. <https://doi.org/10.1080/15305058.2016.1231193>
- Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, 22(2), 154–184. <https://doi.org/10.1080/15305058.2022.2036161>

- Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessments in Education*, 8, Article 5. <https://doi.org/10.1186/s40536-020-00082-1>
- Setzer, J. C., Wise, S., van den Heuvel, J., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, 26(1), 34–49. <https://doi.org/10.1080/08957347.2013.739453>
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response time with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34(3), 213–232. <https://doi.org/10.1111/j.1745-3984.1997.tb00516.x>
- Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record*, 120(12), 1–26. <https://doi.org/10.1177/016146811812001202>
- Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-scale Assessments in Education*, 9(1). <https://doi.org/10.1186/s40536-021-00100-w>
- Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, 32(2), 151–165. <https://doi.org/10.1080/08957347.2019.1577244>
- Wellington, J., Gnamb, T., & Carstensen, C. H. (2024). Identifying disengaged responding in multiple-choice items: Extending a latent class item response model with novel process data indicators. *Educational and Psychological Measurement*, 84(2), 314–339. <https://doi.org/10.1177/00131644231169211>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81. <https://doi.org/10.1006/ceps.1999.1015>
- Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, 28(3), 237–252. <https://doi.org/10.1080/08957347.2015.1042155>
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, 36(4), 52–61. <https://doi.org/10.1111/emip.12165>
- Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, 32(4), 325–336. <https://doi.org/10.1080/08957347.2019.1660350>
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19–38. <http://www.jstor.org/stable/20461807>
- Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343–354. <https://doi.org/10.1080/08957347.2017.1353992>
- Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April 13). *An investigation of motivation filtering in a statewide achievement testing program* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Diego, California, USA.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2
- Wise, S. L., & Ma, L. (2012, duben). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Vancouver, Kanada.

- Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, 22(2), 185–205. <https://doi.org/10.1080/08957340902754650>
- Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, 8(4), 341–351. https://doi.org/10.1207/s15324818ame0804_4
- Zhao, A., Brown, G. T. L., & Meissel, K. (2022). New Zealand students' test-taking motivation: an experimental study examining the effects of stakes. *Assessment in Education: Principles, Policy & Practice*, 29(4), 397–421. <https://doi.org/10.1080/0969594x.2022.2101043>

PhDr. Jiří Štípek, Ph.D. (korespondenční autor)
katedra informačních technologií a technické výchovy
Univerzita Karlova, Pedagogická fakulta
Magdalény Rettigové 4, 116 39 Praha
jiri.stipek@pedf.cuni.cz

prof. RNDr. PhDr. Hana Voňková, Ph.D. et Ph.D.
katedra pedagogiky
Univerzita Karlova, Pedagogická fakulta
Magdalény Rettigové 4, 116 39 Praha
hana.vonkova@pedf.cuni.cz

PhDr. Kateřina Králová, Ph.D.
katedra pedagogiky
Univerzita Karlova, Pedagogická fakulta
Magdalény Rettigové 4, 116 39 Praha
katerina.kralova@pedf.cuni.cz