Shoda hodnotitelů písemného projevu z německého jazyka

Raters´Agreement in the Assessment of Students´ Written Performance in German Language

Soňa Novotná Knotková

Abstrakt: Příspěvek je věnován problematice hodnocení otevřených úloh ve formě písemného projevu z německého jazyka, jenž je ověřován státní maturitní zkouškou. Jsou zde definovány základní pojmy, srovnána analytická a holistická kritéria s přihlédnutím na ověřování jazykových dovedností v cizím jazyce, specifikován písemný projev MZ a představena analytická kritéria a hodnotící škály ordinálního charakteru. Na výzkumném vzorku 1424 učitelů německého jazyka, kteří se účastnili odborné přípravy hodnotitelů v letech 2009–2011, jsou představeny základní metodologicko-statistické přístupy pro zjištění míry shody hodnotitelů a přesnosti a spolehlivosti hodnocení písemného projevu. Analýza a interpretace dílčích výsledků výzkumného šetření přináší cenné informace o problematice hodnocení široce otevřených úloh. Následné závěry a doporučení jsou přínosem také pro praktickou oblast výuky produktivních řečových dovedností v cizích jazycích a jejich hodnocení.

Klíčová slova: hodnocení, shoda hodnotitelů, inter-rater reliabilita, intra-rater reliabilita, hodnotící škála

Abstract: This paper deals with the assessment of open-ended of items in the writing part of the Maturita Exam in German language. The abstract includes definitions of the basic terminology, comparison of analytic and holistic assessment criteria for writing skills in foreign language and specifications for the writing part of the Maturita Exam. Furthermore, analytic criteria and its rating scales of ordinal character are presented. Basic methodological and statistical approaches such as the mean and standard deviation are presented as a tool to measure the raters agreement and the reliability of writing assessment on the sample of 1424 teachers of German language who took part in the rater-training between the years 2009 and 2011. These research data analysis and interpretation serve as a useful source of information about the assess ment of open-ended items. Subsequent conclusions and recommendations can also contribute to the practical sphere of productive skills education in foreign languages and its assessment.

Keywords: assessment, rater agreement, inter-rater reliability, intra-rater reliability, rating scale

1 Úvod

Základním východiskem pro změnu v pojetí, obsahu, cílech a metodách uplatňovaných ve výuce cizích jazyků a pro zcela novou koncepci státní maturitní zkoušky z cizích jazyků byla kurikulární reforma, s níž souvisí potřeba synchronizace strategických směrů a cílů vzdělávací a jazykové politiky České republiky, Evropské unie a Rady Evropy.

Metodické postupy používané ve výuce cizích jazyků a evaluace výsledků vzdělávání v cizích jazycích se na českých školách v posledním desetiletí 20. století výrazně změnily. Postupně se přešlo od lingvistického strukturalismu, gramaticko-překladových metod k výuce orientované na praktické potřeby žáka a akcentující autenticitu výukových materiálů. Jazyk se již nechápe

jako systém izolovaných elementů a struktur, nýbrž jako prostředek sloužící ke komunikaci a k dosažení komunikačního cíle. Toto zcela nové pojetí výuky cizích jazyků se odráží nejen v jazykové police, ale i v evaluačních nástrojích, které ověřují tzv. komplexní jazykovou vybavenost a úroveň jedinců, tj. ověřují nejen receptivní, ale i produktivní řečové dovednosti. Didaktické testy ověřující receptivní řečové dovednosti si v českém pedagogickém prostředí našly již své pevné místo, avšak podstatně méně pozornosti je věnováno právě těm evaluačním nástrojům, které mají ověřit produktivní dovednosti, tj. ústní a písemný projev, dále jejich vzniku, implementaci, objektivitě a reliabilitě.

2 Vymezení základních pojmů

Současnou pedagogickou veřejností je *didaktický test* vnímán jako nástroj, pomocí něhož lze objektivně, spolehlivě a přesně měřit úroveň osvojení vědomostí a dovedností dané skupiny testovaných. Definici didaktického testu jasně a výstižně formuloval P. Byčkovský, a to jako *nástroj systematického zjišťování (měření) výsledků výuky* (Byčkovský, 2007, s. 7; Chráska, 1999, s. 12).

V souvislosti se školskou reformou, kurikulárními změnami a evaluací výsledků českého vzdělávání je kladen důraz na kvalitu didaktických testů, jejich standardizaci, objektivitu, spolehlivost a přesnost jejich vyhodnocování (Byčkovský & Zvára, 2007; Chráska, 1999; Komenda & Mazuchová, 1995). *Hodnocení* testů tvořených uzavřenými úlohami lze považovat za objektivní. Za objektivní lze označit i testy s úlohami požadujícími stručnou odpověď, jejichž posuzování provádí hodnotitelé na základě předem stanoveného klíče správných odpovědí. V případě možného výskytu chyb při posuzování se doporučuje hodnocení otevřených úloh se stručnou odpovědí dvěma, popř. třemi nezávislými hodnotiteli. Problém s hodnocením, jeho objektivitou, spolehlivostí a přesností nastává u testů s široce otevřenými úlohami, tj. *esej-testů* (Byčkovský, 2007, s. 11; Chráska, 2007, s. 188; Chráska, 1999, s. 17). Určitá míra objektivity, spolehlivosti a přesnosti výsledků těchto testů je zpravidla zaručena hodnocením několika nezávislými vyškolenými hodnotiteli, nejčastěji dvěma, jejichž shoda čili konsenzus je celkovým výsledkem daného testu. Pojem *shoda hodnotitelů* (*rater agreement*) označuje míru shody mezi hodnotiteli. Rozlišujeme míru shody mezi dvěma a více hodnotiteli, *inter-rater agreement*, resp. jejich výsledným hodnocením téhož testu neboli výkonu testovaného, a míru shody mezi dvěma či vícero hodnoceními téhož testu, která prováděla tatáž osoba za stejných podmínek, ale v různém čase, *intra-rater agreement*. Poslední zmíněná míra shody je významná z hlediska toho, abychom mohli posoudit, zda je hodnotitel ve svém hodnocení konzistentní. Posuzování či měření shody hodnotitelů je velmi důležité při testování produktivních řečových dovedností, při kterém se uplatňuje subjektivní hledisko hodnotitele, které ovlivňuje výsledek hodnocení. Důležitým předpokladem kvalitního testování je *reliabilita*, tj. spolehlivost a přesnost měření. Na základě počtu hodnotitelů rozlišujeme mezi *inter-rater reliabilitou*, spolehlivost a přesnost hodnocení dvou a více hodnotitelů, a *intra-rater reliabilitou*, spolehlivost a přesnost hodnocení jednoho hodnotitele. Míra shody hodnotitelů, inter-rater reliabilita a intra-rater reliabilita, jsou měřeny různými statistickými výpočty (Byčkovský, 2007; Gavora, 2010; Hendl, 2006; Mareš, 1983; Wirtz & Caspar, 2002). Při volbě vhodné statistické metody u měření reliability hraje velkou roli charakter hodnotících škál. Obvykle se rozlišují tři typy, a to škály nominálního, ordinálního a intervalového charakteru.

Hodnotitelé postupují při hodnocení dle závazných postupů, principů a jednotných kritérií, aby byla zaručena určitá míra objektivity, spolehlivosti a přesnosti. Kritéria hodnocení jakožto nástroj měření výkonu testovaného jsou postavena na principu tzv. *hodnotící škály* (*rater scale*), kdy jednotlivé škály představují důležité aspekty, jevy, charakteristiky či kategorie

hodnocení. Těmto posuzovaným jevům jsou přiřazovány číselné hodnoty. Hovoříme zde o tzv. *škálování* (*rating*), které umožňuje kvantifikovat kvalitativní pojmy (Gavora, 2010, s. 105–120; Průcha, Mareš, & Walterová, 2001, s. 238). Dle způsobu hodnocení rozlišujeme dva základní typy kritérií, a to *analytická* (*analytic criteria*) a *holistická* (*holistic criteria*). Při holistickém hodnocení posuzujeme výkon jedince jako celek, tj., jak testovaný splňuje předem stanovenou celkovou charakteristiku dané hodnotící škály. Při aplikaci analytických kritérií hodnotíme výkon testované osoby odděleně dle jednotlivých dílčích charakteristik, jevů či aspektů. V cizích jazycích jsou hodnotící škály často definovány pomocí tzv. deskriptorů (Společný evropský referenční rámec pro jazyky, 2002; Alderson, Clapham, & Wall, 1995; Shaw & Weir, 2007; Weigle, 2002; Weir, 1990), které přesně vymezují, co by měl testovaný jedinec na určité škále umět. Oba způsoby hodnocení mají své výhody a nevýhody (tabulka 1).

Tabulka 1: Srovnání holistických a analytických kritérií

	holistická kritéria	analytická kritéria
reliabilita	Reliabilita je nižší.	Reliabilita je vyšší.
konstrukční validita	Hodnocení předpokládá, že veškeré posuzované aspekty dané dovednosti jsou na stejné úrovni (na téže hodnotící škále).	Jednotlivé aspekty dané dovednosti jsou posuzovány odděleně a přiřazovány tak různým hodnotícím škálám (testovaný přesně ví, v čem jsou jeho slabé a silné stránky). Kritéria jsou vhodnější pro hodnocení jazykových dovedností v cizím jazyce.
praktičnost	Hodnocení není časově náročné, ale jsou kladeny větší nároky na odbornou přípravu a zkušenosti hodnotitele.	Hodnocení je časově a ekonomicky náročné.
dopad/vliv	Obecný popis výsledku neposkytuje dostatečnou zpětnou vazbu o úrovni osvojení vědomostí a dovedností.	Kritéria mají diagnostický význam, poskytují zpětnou vazbu o úrovni osvojení vědomostí a dovedností.
autentičnost	Holistické hodnocení je přirozenější proces než analytické hodnocení.	Hrozí riziko opomíjení analytických kritérií a jejich přizpůsobení se holistickému hodnocení.

Pozn. Upraveno podle Shaw (2007, s. 153) a Weigle (2002, s. 121).

Volba použití holistických nebo analytických kritérií záleží na řadě faktorů, např. na tom, co chceme přesně testováním ověřit, kolik bude testovaných, kdo a jakým způsobem bude hodnocení provádět, k jakému účelu mají výsledky testování sloužit, v jakém čase mají být testy vyhodnoceny, jaké máme finanční, časové, organizační a personální možnosti pro odbornou přípravu hodnotitelů.

3 Postavení písemného projevu v maturitní zkoušce, jeho specifikace a hodnocení

Písemný projev je jedna ze čtyř základních jazykových dovedností, která se spolu s dalšími řečovými dovednostmi, jako je ústní projev, poslech a čtení s porozuměním, ověřuje v rámci

státní maturitní zkoušky z cizích jazyků. Komplexní maturitní zkouška je připravována *Centrem pro zjišťování výsledků vzdělávání* (CERMAT) ve dvou úrovních obtížnosti, základní a vyšší. Písemná práce se skládá ze dvou částí, z nichž každá část ověřuje na základě zadání dva různé slohové útvary, které se od sebe liší tematickým zaměřením a komunikační situací. Čas vymezený na splnění zadání písemné práce je pro základní úroveň obtížnosti 60 minut, pro vyšší 90 minut. Požadovaný rozsah písemné práce je u základní úrovně obtížnosti v 1. části 120–150 slov, ve 2. části 60–70 slov, u vyšší úrovně v 1. části 210–240 slov, ve 2. části 100–120 slov. Testované osoby své odpovědi zapisují do záznamového archu, jehož naskenovaná kopie slouží hodnotitelům jako podklad pro hodnocení. Celkový počet bodů, který může žák za splnění zadání obou částí získat, je 36, za 1. část 24 bodů a za 2. část 12 bodů.

Základními východisky při rozhodování o koncepci kritérií a způsobu hodnocení byly poznatky uváděné v zahraničních publikacích (Alderson, 1995; Arter & Chappuis, 2007; Bachmann & Palmer, 1996; McNamara, 2000; Shaw & Weir, 2007; Weigle, 2002) a zkušenosti zahraničních kolegů z oblasti evaluace výsledků vzdělávání v cizích jazycích. S ohledem na jejich poznatky a zkušenosti se při hodnocení písemného i ústního projevu používají analytická kritéria, hodnotící škála ordinálního charakteru. Vzhledem k tomu, že je písemná práce tvořena dvěma částmi, v nichž se ověřují dva různé typy textů v různé délce, byla vyvinuta dvě velmi podobná analytická kritéria. Kritéria pro 1. část (dlouhý písemný projev) jsou tvořena 8 oddíly. V oddíle I posuzujeme, zda žák splnil zadání z hlediska požadované charakteristiky textu, zmínil body strukturovaného zadání a splnil požadovanou délku písemného projevu; v oddíle II hodnotíme, jak žák myšlenkově, po stránce kvality obsahu, splnil zadání; v oddíle III koherenci textu; v oddíle IV správnost a rozsah použitých kohezních prostředků; v oddíle V přesnost slovní zásoby, tj. počet chyb; v oddíle VI rozsah použité slovní zásoby; v oddíle VII přesnost mluvnických prostředků, tj. počet chyb; v oddíle VIII rozsah mluvnických prostředků. Každý oddíl je tvořen několika hierarchicky uspořádanými deskriptory, které charakterizují jednotlivé dílčí posuzované aspekty či jevy a odpovídají bodové škále od 0 do 3 bodů. Za každý oddíl může testovaný získat max. 3 body, za celou 1. část tedy celkem 24 bodů. Kritéria pro 2. část (krátký písemný projev) jsou konstruována na stejném principu. S ohledem na požadovaný rozsah práce jsou tvořena 4 oddíly, přičemž za každý oddíl může žák získat max. 3 body, za celou 2. část tedy celkem 12 bodů. Celkem za obě dvě části to činí 36 bodů.

Dle koncepce, která platila do roku 2010, měly být písemné a ústní projevy hodnoceny na školách dvěma nezávislými vyškolenými a certifikovanými hodnotiteli podle jednotných kritérií a principů. Shoda neboli konsenzus dílčích hodnocení měl být celkovým hodnocením výkonu testovaného. Již v roce 2011 nebylo hodnocení dvěma nezávislými hodnotiteli u písemného projevu realizováno a práce na školách hodnotil jen jeden certifikovaný učitel. Na podzim roku 2011 došlo ke změně koncepce hodnocení a písemný projev je vyhodnocován centrálně několika vyškolenými externími hodnotiteli. Hodnocení a zkoušení ústního projevu bylo ponecháno v kompetenci škol, tj. v kompetenci vyškolených a certifikovaných učitelů.

Pro dosažení co nejvyšší možné míry objektivity, spolehlivosti, přesnosti a srovnatelnosti hodnocení písemného projevu byla ze strany CERMATu učiněna jistá opatření. Jednotné zadání je strukturované, tzn., že struktura zadání přesně vymezuje požadavky na jeho splnění, a to formou tzv. dílčích bodů zadání. Ke každému zadání je vytvořen tzv. metodický balíček, jehož součástí jsou přesně definované požadavky na výkon žáka vhledem k ověřované referenční jazykové úrovni a ke znění zadání, dále několik ukázek prací žáků z procesu pretestování a detailní komentáře k jejich hodnocení. Důležitým předpokladem pro objektivitu a reliabilitu je kvalitní a efektivní odborná příprava hodnotitelů, jejíž součástí jsou výborné znalosti a implementace referenčních jazykových úrovní dle *Společného evropského*

referenčního rámce pro jazyky, dodržení striktně stanovených jednotných principů a postupů při hodnocení a zajištění stejného sémantického chápaní jednotlivých deskriptorů.

4 Výzkumné šetření

V souvislosti s původně navrženou koncepcí hodnocení písemného projevu na školách probíhala v letech 2009–2011 odborná příprava hodnotitelů písemného projevu. Školení bylo realizováno e-learningovou a prezenční formou. E-learningová část se skládala z pěti modulů zakončených povinným on-line testem. Na studium bylo vymezeno 16,5 hodin. Po úspěšném absolvování e-learningové části následovala prezenční část s osmihodinovou dotací, která se skládala ze dvou částí a prověřovala na základě hodnocení několika písemných prací teoretické znalosti získané on-line studiem. V závěru prezenčního studia vykonali účastníci tzv. certifikační řízení, jehož cílem bylo účastníkům za splnění jistých podmínek udělit certifikát opravňující k výkonu funkce hodnotitele. Při certifikaci posuzovali hodnotitelé dvě práce (dlouhý a krátký písemný projev) na základní úrovni obtížnosti a dvě práce stejného formátu na vyšší úrovni obtížnosti. Certifikační řízení týkající se vyšší úrovně obtížnosti nebylo povinné, účastnili se ho jen ti učitelé, pro jejichž pedagogickou praxi bylo získání certifikátu opravňujícího hodnotit vyšší úroveň nezbytné. Základním kritériem pro úspěšné absolvování školení a získání certifikátu byla určitá míra shody hodnotitele s expertním hodnocením. Některé ze základních poznatků týkajících se míry shody hodnotitelů a reliability procesu hodnocení se nyní pokusíme stručně shrnout v rámci dílčího výzkumného šetření.

Odborné přípravy se zúčastnilo celkem 3 149 učitelů německého jazyka, z nichž 3 044 jedinců se rozhodlo pro certifikaci vyšší úrovně. Při certifikačním řízení týkající se hodnocení prací (dlouhý a krátký písemný projev) základní úrovně obtížnosti neuspělo 3,6 % uchazečů, u prací vyšší úrovně neuspělo 10,4 % uchazečů. Do výzkumného šetření byli zahrnuti jen ti jedinci, u nichž byla k dispozici data za hodnocení práce 1. části písemného projevu základní i vyšší úrovně obtížnosti, která byla součástí zadání tzv. domácího úkolu po 1. části prezenčního školení, a z hodnocení práce 1. části písemného projevu základní i vyšší úrovně obtížnosti, která byla určena k certifikačnímu řízení. Hlavním důvodem pro výběr těchto jedinců a dat je možnost zkoumání, zda se míra shody hodnotitelů a inter-rater reliabilita školením zvýšila, zda se učitelé ve svém hodnocení přiblížili k hodnotám uděleným expertním týmem, a zda tak byl proces odborné přípravy v souvislosti s jistými omezeními a v určitých měřítkách efektivní a směřoval ke stanoveným cílům a účelům.

Výzkumné šetření se týkalo 1424 učitelů německého jazyka, z nichž se 604 jedinců účastnilo certifikace označené CI, 760 jedinců certifikace CII. Všichni učitelé hodnotili stejnou práci, která byla určena za domácí úkol (dále jen DU) mezi 1. a 2. částí prezenčního školení. Nejvíce respondentů bylo z řad učitelů, kteří vyučovali na středních odborných školách (tabulka 2). U certifikačního řízení bylo 66 respondentů neúspěšných, tj. 4,6 % z celkového počtu 1424.

Tabulka 2: Výzkumný vzorek - hodnotitelé dle typu školy

	G	G+SOŠ	SOŠ	SOŠ+SOU	SOU	VOŠ a/nebo VŠ	VOŠ+SOŠ	jiné	celkem
DU	422	58	**481**	290	54	1	101	17	**1424**
CI	189	26	**222**	154	14	1	52	6	664
CII	233	32	**259**	136	40	0	49	11	760

Pozn. DU – domácí úkol, CI – certifikace, CII – certifikace.

Dle uvedených základních výpočtů (tabulka 3), průměru a směrodatné odchylky hodnot čili bodů za jednotlivé oddíly a za celou opravenou písemnou práci 1. části DU, CI a CII základní úrovně obtížnosti můžeme posoudit, nakolik se hodnotitelé lišili od hodnot/bodů udělených expertním týmem, ve kterých oddílech dosahovali hodnotitelé nejvyšší/nejnižší shody a zda se hodnotitelé ve svém hodnocení u certifikačního řízení zlepšili, tj. zda byla jejich míra shody vyšší než u práce, která byla zadána jako domácí úkol v průběhu školení. V hodnocení práce DU a CII byli učitelé mírnější než expertní tým. U práce CI bylo celkové hodnocení téměř totožné s hodnocením expertního týmu. Jedny z nejvyšších hodnot rozptylu bodů se u všech prací vyskytují v oddíle VII a VIII, tj. v hodnocení chyb a rozsahu použitých mluvnických prostředků. Na základě hodnot směrodatné odchylky za celou práci jak u základní, tak i u vyšší úrovně obtížnosti (srov. tabulky 3 a 6), lze konstatovat, že se míra shody učitelů během školení zvýšila.

Tabulka 3: Průměr a směrodatná odchylka hodnot/bodů za jednotlivé oddíly kritérií a za celou práci DU, CI, CII zkladní úrovně obtížnosti

oddíly	I	II	III	IV	V	VI	VII	VIII	celkem
DU									
E	3	2	2	2	2	2	2	2	17
průměr	2,76	2,14	2,14	1,90	2,35	1,90	2,05	1,99	17,23
směr. odch.	0,45	0,49	0,46	0,42	0,55	0,46	0,50	0,53	**2,18**
CI									
E	2	2	2	1	2	2	2	1	14
průměr	1,92	1,80	2,16	1,33	1,75	1,77	1,82	1,42	13,97
směr. odch.	0,40	0,43	0,43	0,53	0,45	0,45	0,48	0,51	**1,73**
CII									
E	3	2	2	1	2	2	2	1	15
průměr	2,65	2,23	2,06	1,24	1,98	1,79	1,90	1,56	15,40
směr. odch.	0,48	0,49	0,39	0,48	0,36	0,47	0,50	0,54	**1,80**

Pozn. DU – domácí úkol; E – expertní tým; CI – certifikace, CII – certifikace.

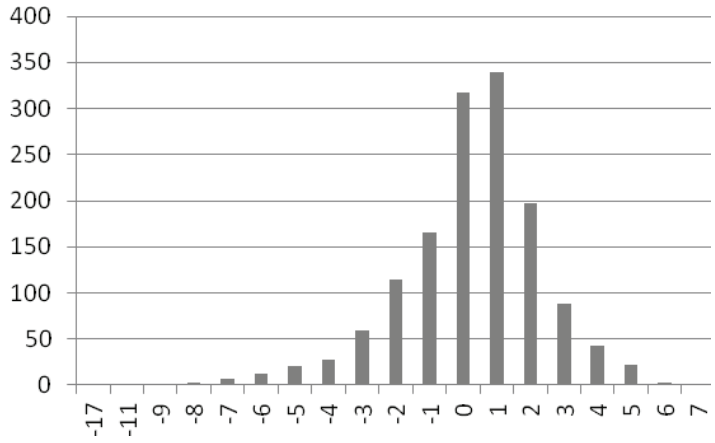
Na základě průměru vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem u jednotlivých oddílů kritérií a za celou práci 1. části DU, CI a CII základní úrovně obtížnosti (tabulka 4) lze posuzovat, ve které práci a ve kterých oddílech byli učitelé při svém hodnocení mírnější/přísnější než expertní tým. U práce DU a CII posuzovali hodnotitelé oddíl I kritérií (splnění zadání) daleko přísněji než expertní tým. Přísnější byli učitelé také u hodnocení oddílu VI (rozsahu slovní zásoby). Oddíl VIII (rozsah mluvnických prostředků) u práce CI a CII hodnotiteli učitelé daleko mírněji než expertní tým.

Tabulka 4: Průměr vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za jednotlivé oddíly kritérií a za celou práci DU, CI, CII zkladní úrovně obtížnosti

oddíly	I	II	III	IV	V	VI	VII	VIII	celkem
DU									
průměr	**-0,24**	0,14	0,14	-0,10	0,35	-0,10	0,05	-0,01	**0,23**
CI									
průměr	-0,08	-0,20	0,16	0,33	-0,25	**-0,23**	-0,18	**0,42**	-0,03
CII									
průměr	**-0,35**	0,23	0,06	0,24	-0,02	**-0,21**	-0,10	**0,56**	**0,40**

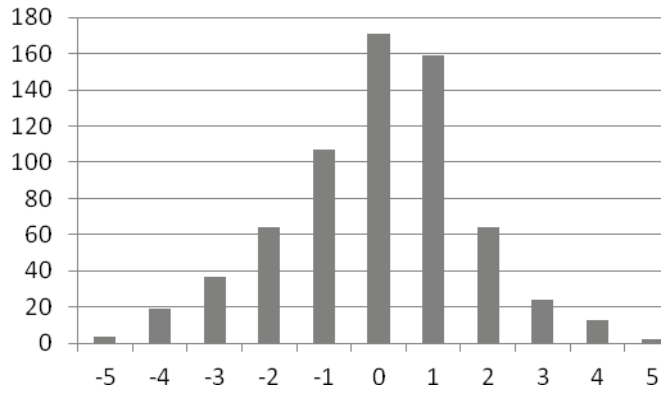
Grafy 1, 2 a 3 zobrazují četnosti vzdáleností hodnot/bodů za celou práci DU, CI a CII udělených učiteli a expertním týmem, tj. kolik hodnotitelů se lišilo od expertního týmu o 1 a více bodů nad/pod hodnotou udělenou expertním týmem za celou práci. Většina hodnotitelů udělovala stejný počet bodů za celou práci jako expertní tým (vyjma práce DU). U práce DU a CI byli učitelé při svém posuzování mírnější než expertní tým. Rozptyl hodnot byl největší při nácviku, tj. u práce, která byla hodnocena jako DU v průběhu školení. Při hodnocení certifikačních prací se rozptyl hodnot výrazně zmírnil.

Graf 1: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za celou práci DU základní úrovně obtížnosti



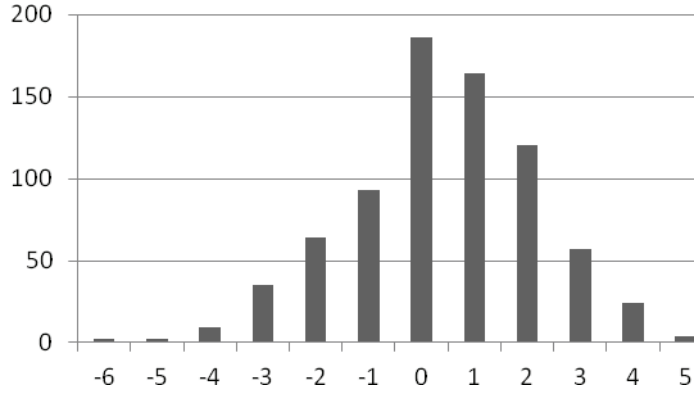
Pozn. osa x – vzdálenost hodnot/bodů od expertního týmu za celou práci DU, osa y – počet učitelů.

Graf 2: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za celou práci CI základní úrovně obtížnosti



Pozn. osa x – vzdálenost hodnot/bodů od expertního týmu za celou práci CI, osa y – počet učitelů.

Graf 3: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za celou práci CII základní úrovně obtížnosti



Pozn. osa x – vzdálenost hodnot/bodů od expertního týmu za celou práci CII, osa y – počet učitelů.

Dle četnosti vzdáleností hodnot/bodů mezi učiteli a expertním týmem za jednotlivé oddíly (tabulka 5) se většina hodnotitelů v jednotlivých oddílech, vyjma oddílu VIII (rozsah mluvnických prostředků) u práce CII, shoduje s bodovým hodnocením expertního týmu. Většina hodnotitelů, vyjma třech jedinců v oddíle I u práce DU, se neliší v žádném oddíle o tři body než expertní tým.

Tabulka 5: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za jednotlivé oddíly práce DU, CI, CII základní úrovně obtížnosti

oddíly	I	II	III	IV	V	VI	VII	VIII
DU								
-3	3	0	0	0	0	0	0	0
-2	7	2	4	1	5	4	4	2
-1	317	83	53	203	37	217	132	203
0	**1097**	**1056**	**1102**	**1157**	**836**	**1119**	**1072**	**1030**
1	0	283	265	63	546	84	216	189
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
CI								
-3	0	0	0	0	0	0	0	0
-2	0	0	0	0	0	2	0	0
-1	83	140	17	16	172	154	146	3
0	**553**	**515**	**523**	**419**	**489**	**500**	**491**	**379**
1	28	9	124	226	3	8	27	281
2	0	0	0	3	0	0	0	1
3	0	0	0	0	0	0	0	0
CII								
-3	0	0	0	0	0	0	0	0
-2	1	1	0	0	0	3	0	0
-1	263	20	38	16	56	173	136	5
0	**496**	**545**	**639**	**547**	**664**	**567**	**566**	332
1	0	194	83	196	40	17	58	**412**
2	0	0	0	1	0	0	0	11
3	0	0	0	0	0	0	0	0

Následující část této kapitoly bude věnována interpretaci výsledků týkajících se hodnocení prací vyšší úrovně obtížnosti.

Dle hodnot uvedených v tabulce 6 se hodnotitelé v posuzování práce CI a CII opět zlepšili, jejich míra shody je vyšší, tak jako u základní úrovně obtížnosti. V porovnání s hodnotami z tabulky 3 základní úrovně je však rozptyl hodnot/bodů u vyšší úrovně větší, tzn., že míra shody hodnotitelů je u této úrovně nižší než míra shody učitelů u základní úrovně. Ve většině oddílů (vyjma oddílu I u práce CII, oddílu IV u práce VI a CII, oddílu VII u práce I a oddílu VIII práce CII) byla hodnota směrodatné odchylky větší než u základní úrovně. U hodnocení

práce DU byli hodnotitelé přísnější než expertní tým, ale u posuzování práce CI a CII naopak mírnější.

Tabulka 6: Průměr a směrodatná odchylka hodnot/bodů za jednotlivé oddíly kritérií a za celou práci DU, CI, CII vyšší úrovně obtížnosti

oddíly	I	II	III	IV	V	VI	VII	VIII	celkem
DU									
E	2	2	2	2	2	2	2	2	16
průměr	2,22	1,65	1,83	1,64	1,74	1,80	1,85	1,77	14,50
směr. odch.	0,59	0,62	0,49	0,55	0,56	0,54	0,56	0,55	**2,67**
CI									
E	2	1	1	2	2	2	1	1	12
průměr	2,20	1,58	1,45	1,32	1,78	1,59	1,80	1,56	13,29
směr. odch.	0,45	0,53	0,52	0,49	0,48	0,54	0,47	0,51	**1,99**
CII									
E	2	2	3	2	2	1	2	2	16
průměr	2,12	1,84	2,47	2,11	2,23	2,08	2,28	2,04	17,16
směr. odch.	0,38	0,50	0,54	0,39	0,45	0,47	0,53	0,53	**1,88**

Pozn. DU – domácí úkol; E – expertní tým; CI – certifikace, CII – certifikace.

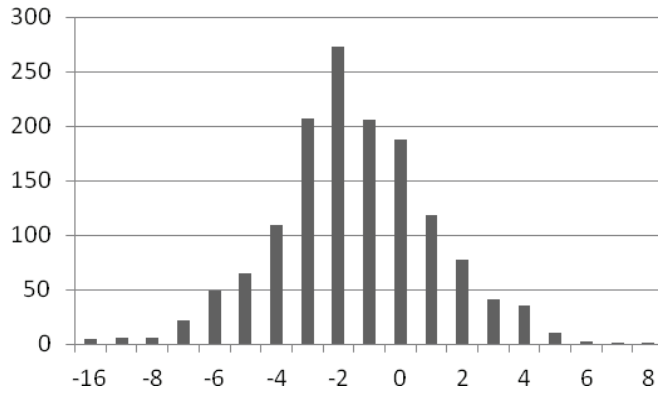
Tabulka 7 poskytuje informace o tom, ve kterých oddílech kritérií a při které práci byli učitelé přísnější/mírnější než expertní tým. Většinu oddílů u práce DU hodnotiteli učitelé přísněji než expertní tým (vyjma oddílu I). U hodnocení práce CI byli hodnotitelé daleko přísnější v oddíle IV (kohezní prostředky) než expertní tým, naopak v oddíle VII (chyby v mluvnických prostředcích) daleko mírnější. Oddíl III (koherence textu) u práce CII byl učiteli hodnocen přísněji než expertní tým, naopak oddíl VI (rozsah slovní zásoby) daleko mírněji. Obě práce u certifikace posuzovali hodnotitelé mírněji než expertní tým. Práci DU hodnotili učitelé přísněji než expertní tým.

Tabulka 7: Průměr vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za jednotlivé oddíly kritérií a za celou práci DU, CI, CII vyšší úrovně obtížnosti

oddíly	I	II	III	IV	V	VI	VII	VIII	celkem
DU									
průměr	**0,22**	-0,35	-0,17	-0,36	-0,26	-0,20	-0,15	-0,23	-1,50
CI									
průměr	0,20	0,58	0,45	**-0,68**	-0,22	-0,41	**0,80**	0,56	1,29
CII									
průměr	0,12	-0,16	**-0,53**	0,11	0,23	**1,08**	0,28	0,04	1,16

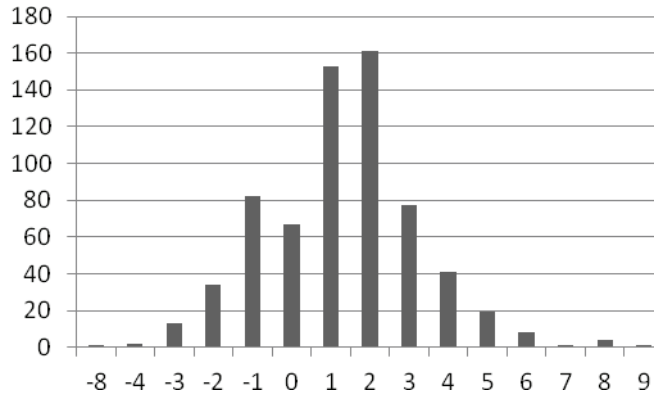
Srovnáme-li grafy 1–6 o četnosti vzdáleností hodnot/bodů za celou práci DU, CI a CII obou úrovní obtížnosti udělených učiteli a expertním týmem, docházíme k závěru, že míra shody hodnotitelů s expertním týmem u základní úrovně je vyšší než u vyšší úrovně. Rozptyl hodnot je menší. Při hodnocení práce DU vyšší úrovně byli učitelé přísnější než expertní tým, u práce CI a CII byli naopak mírnější. V průběhu školení byl rozptyl hodnot za celou práci větší než rozptyl těchto hodnot u certifikačního řízení.

Graf 4: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za celou práci DU vyšší úrovně obtížnosti



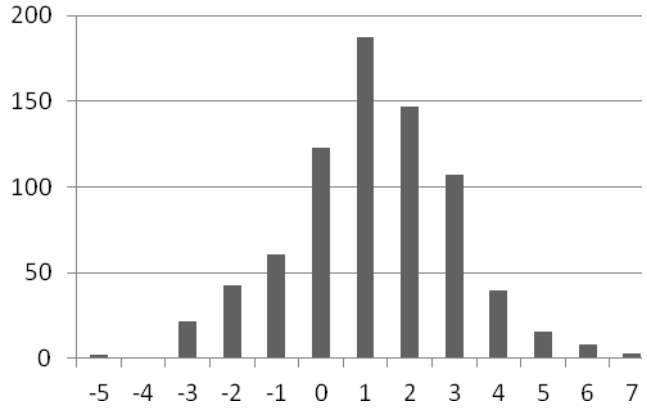
Pozn. osa x – vzdálenost hodnot/bodů od expertního týmu za celou práci DU, osa y – počet učitelů.

Graf 5: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za celou práci CI vyšší úrovně obtížnosti



Pozn. osa x – vzdálenost hodnot/bodů od expertního týmu za celou práci CI, osa y – počet učitelů.

Graf 6: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za celou práci CII vyšší úrovně obtížnosti



Pozn. osa x – vzdálenost hodnot/bodů od expertního týmu za celou práci CI, osa y – počet učitelů.

Na základě hodnot v tabulce 8 o četnosti vzdáleností hodnot/bodů mezi učiteli a expertním týmem za jednotlivé oddíly můžeme konstatovat, že odborná příprava hodnotitelů byla u vyšší úrovně méně uspokojivá než u základní úrovně, jelikož počet učitelů, kteří se od expertního posouzení liší o dva body v jednotlivých oddílech, je vyšší než u základní úrovně (srov. tabulky 5 a 8). V oddílech II (kvalita splnění zadání), IV (kohezní prostředky), VII a VIII (chyby a rozsah mluvnických prostředků) u práce CI udělovala většina učitelů o jeden bod více než expertní tým, u práce CII tato situace nastala v oddíle VI (rozsah slovní zásoby). Ani jeden hodnotitel se nelišil v žádném oddíle od expertního týmu o tři body.

Tabulka 8: Četnost vzdáleností hodnot/bodů mezi hodnotiteli a expertním týmem za jednotlivé oddíly práce DU, CI, CII vyšší úrovně obtížnosti

oddíly	I	II	III	IV	V	VI	VII	VIII
DU								
-3	0	0	0	0	0	0	0	0
-2	6	14	6	11	6	5	11	12
-1	107	571	297	527	447	364	310	374
0	**880**	**740**	**1059**	**848**	**889**	**962**	**982**	**962**
1	431	99	62	38	82	93	121	76
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
CI								
-3	0	0	0	0	0	0	0	0
-2	0	0	0	5	0	2	0	0
1	14	1	3	**441**	166	279	1	1
0	**504**	288	**360**	216	**480**	**369**	150	294
1	146	**363**	297	2	18	14	**495**	**365**
2	0	12	4	0	0	0	18	4
3	0	0	0	0	0	0	0	0
CII								
-3	0	0	0	0	0	0	0	0
-2	0	0	17	0	0	0	0	0
-1	13	163	371	20	9	1	31	91
0	**642**	**554**	**372**	**636**	**567**	53	**487**	**548**
1	105	43	0	104	184	**594**	242	121
2	0	0	0	0	0	112	0	0
3	0	0	0	0	0	0	0	0

5 Závěr

Na základě interpretace výsledků dílčího výzkumného šeření vztahujícího se na zkoumání míry shody hodnotitelů a spolehlivosti a přesnosti hodnocení písemného projevu z německého jazyka lze s ohledem na hodnocení konkrétních prací konstatovat, že se hodnotitelé v průběhu nácviku posuzování těchto široce otevřených úloh zlepšili a jejich míra shody mezi nimi i s expertním posouzením je vyšší. V hodnocení prací základní i vyšší úrovně obtížnosti jsou

hodnotitelé mírnější než expertní tým. Míra shody hodnotitelů je u základní úrovně vyšší než u vyšší úrovně obtížnosti. Lze se domnívat, že tento rozdíl je způsoben mírou znalostí a implementací jazykových úrovní dle *Společného evropského referenčního rámce pro jazyky* a mnohdy i samotnou jazykovou vybaveností učitelů německého jazyka. Nejnižší míra shody se vyskytuje v oddílech ověřující mluvnické prostředky a slovní zásobu. Tuto nízkou míru shody v těchto oddílech nelze označit za negativní, jelikož je mnohdy velmi obtížné rozlišovat, zda je daná chyba lexikálního či morfologického charakteru, popř. i kohezního. Důležité je, zda je tato chyba identifikována a v některém z oddílu (slovní zásoby, mluvnických či kohezních prostředků) penalizována. Zřetel je brán především na míru shody za celou práci. Zdůraznit můžeme i to, že se hodnotitelé (vyjma třech jedinců při hodnocení práce DU základní úrovně obtížnosti) neliší od expertního týmu v žádném oddíle kritérií o tři body. Dle uvedených výsledků a závěrů můžeme konstatovat, že koncepce odborné přípravy hodnotitelů splňuje v souvislosti s jistými omezeními a v určitých aspektech očekávání a svůj záměr. Kriticky lze však pohlížet na praktickou část prezenčního školení, jeho hodinovou dotaci, malý počet ohodnocených prací jak v průběhu školení, tak i u certifikačního procesu. Za negativní můžeme považovat i tu skutečnost, že nebyla v průběhu školení důsledně monitorována kalibrace a konzistentnost hodnocení jednotlivých učitelů.

Hodnocení otevřených úloh se širokou odpovědí přináší mnoho úskalí. Především vysoká míra objektivity a spolehlivosti skórování není vzhledem k uplatňování subjektivních hledisek hodnotitelů zcela zaručena. Určitou míru objektivity, spolehlivosti a přesnosti těchto úloh lze zajistit jejich konstrukcí, vhodným výběrem kritérií hodnocení a vysokou mírou kompetentnosti hodnotitelů k jejich posuzování. Zejména odborná příprava hodnotitelů musí být velmi kvalitní a efektivní, i když ta je většinou podmíněna ekonomickými, časovými, organizačními a personálními možnostmi. Vzhledem ke všem problémům, které tyto úlohy s sebou přináší, nelze tyto v běžném procesu vzdělávání a evaluace výsledků vzdělávání opomíjet, zejména tehdy, když chceme ověřit komplexní dovednosti žáků, a nezařazovat tak do běžného testování, ať už na úrovni školy či státu. Právě díky jejich komplexnímu charakteru získáme cenné informace nejen o výkonu testovaného jedince, ale i o výsledcích své pedagogické práce.

Literatura

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Arter, J., & Chappuis, J. (2007). *Creating & recognizing quality rubrics*. New Jersey: Upper Saddle River.

Bachmann, L., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Byčkovský, P., & Zvára, K. (2007). *Konstrukce a analýza testů pro přijímací řízení*. Praha: Pedagogická fakulta UK.

Byčkovský, P. (1982). *Základy měření výsledků výuky. Tvorba didaktického testu*. Praha: ČVUT.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. (2001). Cambridge: Cambridge University Press.

Gavora, P. (2010). *Úvod do pedagogického výzkumu*. Brno: Paido.

Hendl, J. (2009). *Přehled statistických metod*. *Analýza a metaanalýza dat*. Praha: Portál.

Chráska, M. (2007). *Metody pedagogického výzkumu. Základy kvantitativního výzkumu*. Praha: Grada Publishing.

Chráska, M. (1999). *Didaktické testy*. *Příručka pro učitele a studenty učitelství*. Brno: Paido.

Komenda, S., & Mazuchová, J. (1995). *Tvorba a testování testu*. Olomouc: UP.
Mareš, J. (1983). Jak zjišťovat reliabilitu pozorování? *Pedagogika*, *33*(2), 169–189.
McNamara, T. (1996). *Measuring second language performance*. Londýn: Longman.
McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.
Milanovic, M. (Ed.). (1998). *Multilingual glossary of language testing terms*. Cambridge: Cambridge University Press.
Průcha, J., Walterová, E. & Mareš, J. (2001). *Pedagogický slovník*. Praha: Portál.
Rada Evropy (2002). *Společný evropský referenční rámec pro jazyky*. Olomouc: UP.
Shaw, S. D., & Weir, C. J. (2007). *Examing writing. Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Weir, C. J. (1990). *Communicative language testing*. Hemel Hempstead: Prentice Hall.
Weir, C. J. (1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall.
Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Kontakt

Mgr. Soňa Novotná Knotková
Centrum pro zjišťování výsledků vzdělávání (CERMAT)
Jankovcova 933/63
170 00 Praha 7
knotkova@cermat.cz