Jak zjišťovat reliabilitu pozorování?

PhDr. JIŘÍ MAREŠ, CSc.,
kabinet pedagogiky lékařské fakulty UK, Hradec Králové

ÚVOD

Marxistická pedagogika se dnes rozvíjí v obou základních směrech: jako svébytný vědní obor i jako aplikovaná disciplína. Přitom v nedávném období se kladl důraz na obsahovou stránku a aktuálnost úkolů, jež měla pedagogika řešit; méně pozornosti se věnovalo interpretaci získaných výsledků a korektnosti jejich zobecňování.[1])

V poslední době však vzrůstá zájem o řešení metodických problémů, které vznikají jednak při převádění teoretických výsledků do pedagogické praxe, jednak při teoretickém zpracování praktických zkušeností. Jednotlivé problémové okruhy nejsou ovšem propracovány stejně. Vezměme např. metodologické problémy pozorování. Pozorování jako metoda patří k nejdůležitějším přirozeným zdrojům primárních údajů o pedagogickém procesu.[2]) Skutečnost pozorují učitelé, žáci, ředitelé, inspektoři i výzkumní pracovníci. Pozorování je základem běžných výzkumů na školách, ale též náslechů, hospitací a inspekcí. Přesto se v pedagogice až donedávna příliš nezkoumalo, nakolik jsou výsledky pozorování zatíženy chybami.

Cílem naší práce je seznámit odbornou veřejnost s pokroky, kterých bylo dosaženo při zkoumání reliability standardizovaného pozorování. Jinak řečeno, pokusíme se zodpovědět dvě otázky: na čem závisí přesnost, spolehlivost, stálost, konzistentnost (neboli reliabilita) údajů získaných pozorováním, jakými statistickými postupy můžeme spolehlivost pozorování spočítat.

Zdánlivě jde o otázky ryze teoretické, které spadají do obecné metodologie pedagogiky. Zároveň však správná odpověď na ně má velmi praktické důsledky. Říká nám, nakolik se můžeme na získané výsledky spolehnout, nakolik jsme oprávněni je zobecnit.

Reliabilita má z metodologického hlediska některé zajímavé charakteristiky. Především je nutnou, nikoli však postačující podmínkou kvalitního výzkumu a seriózního zobecnění dat. Vysoká reliabilita sama o sobě není však ještě zárukou dobrých vědeckých výsledků, ale dobrých vědeckých výsledků nemůžeme dosáhnout nespolehlivými postupy, postupy s nízkou reliabilitou.[3])

Na rozdíl od validity, platnosti (jež úzce souvisí s filozofickou a světonázorovou orientací autora, zastávanou pedagogickou koncepcí atd.) je dosažení uspokojivé reliability do jisté míry věc technická — záležitost statistické povahy. Protože v socialistických zemích nebyly zatím otázky reliability pozorování podrobnější studovány [4]), můžeme — obdobně, jako to činí sovětští autoři[5]) — z nemarxistické pedagogiky převzít a rozvinout řadu matematických postupů, které nejsou přísně vázány na konkrétní pedagogické kategorie a vyznačují se velkým stupněm univerzálnosti.

VYMEZENÍ ZÁKLADNÍCH POJMŮ

Dříve než přistoupíme k rozboru reliability pozorování, musíme vymezit některé pojmy, především standardizované pojmy pozorování a pozorovací technika.

Standardizované pozorování je činnost pozorovatele spočívající v záměrném, cílevědomém, systematickém a relativně objektivním sledování smyslově vnímatelných jevů, které nebyly vyvolány zásahem pozorovatele. Proces pozorování v širším pojetí zahrnuje vnímání, třídění, záznam, hodnocení a interpretaci údajů o pozorovaném objektu. Činnost pozorovatele je díky standardizované pozorovací technice do jisté míry formalizována.[6])

Pozorovací technika je objektivovaný návod (tedy prostředek, nástroj) pro pozorování[7]) předpisující pozorovatelovu činnost tak, aby bylo možné získat reliabilní a validní údaje o pozorovaném objektu. Určuje účel i podmínky pozorování, sledované objekty, vybírá a definuje vhodné vlastnosti (stavy, znaky) pozorovaných objektů, stanovuje způsob jejich sledování, popisu a rozboru; předpisuje způsob záznamu, zpracování a vyhodnocení získaných dat. K základním typům standardizovaných pozorovacích technik patří techniky škálovací, znakové a kategoriální.

Můžeme je charakterizovat takto:

Posuzovací škála je pozorovací technika umožňující globálně posoudit vlastnosti sledovaného objektu až po skončení pozorování. Obvykle vyžaduje, aby pozorovatel zhodnotil sledovaný znak (resp. znaky) z kvalitativního hlediska a výsledek zaznamenal na jisté škále, často kvantitativní.

Příklad[8]):

Chování učitele vůči žákům bylo

nejisté	1	2	3	4	5	6	7	rozhodné
tvrdé	1	2	3	4	5	6	7	laskavé

Znakový systém je pozorovací technika umožňující zachytit výskyt vybraných znaků sledovaného objektu ve stanovených časových úsecích. Každému znaku je přidělen určitý symbol, jenž indikuje pouze přítomnost či nepřítomnost znaku v jistém časovém úseku bez ohledu na to, kolikrát se vyskytl či v jaké intenzitě se vyskytl.

Příklad[9]):

I.	II.	III.	IV.	V.	VI.	Sledovaný znak
						1. Učitel na sebe soustředil pozornost žáků
						2. Učitel soustředil pozornost na žáka
						10. Učitel povzbuzoval žáky, aby volně vyjádřili své názory

Kategoriální systém je pozorovací technika umožňující zachytit vybrané znaky sledovaného objektu pokaždé, kdykoli se vyskytnou, a v pořadí, v jakém se vyskytly. Každému znaku je přidělen určitý kód (nejčastěji numerický či alfanumerický) a pozorovatel kóduje průběh sledovaného jevu v předem určených pravidelných intervalech. Výsledkem jsou údaje o četnosti výskytu sledovaných znaků a o jejich posloupnosti.

Příklad[10])

Je dán systém deseti kategorií:

Učitel 1. Akceptuje žákovy pocity. Konstruktivním způsobem projevuje sympatie.
2. Chválí a povzbuzuje. Souhlasí s žákovým výkonem.
3. Akceptuje nebo používá žákovy názory a myšlenky.
4. Klade otázky.
5. Vykládá.
6. Dává pokyny či příkazy.
7. Kritizuje žákovu činnost nebo chování.

Žák 8. Odpovídá učiteli až po vyvolání, bez iniciativy.
9. Spontánní, iniciativní odpověď učiteli.
10. Ticho nebo zmatek.

Pozorovatel simultánně kóduje průběh vyučovací hodiny podle předepsaných pravidel v třísekundových intervalech. Výsledkem záznamu je posloupnost numerických kódů zastupujících výše uvedené kategorie činnosti, např.

0006070644068825555 atd.

Tato posloupnost kódů se potom přepisuje do matice 10 × 10 a dále zpracovává.

Všechny tři výše uvedené pozorovací techniky můžeme charakterizovat i jinak. Z hlediska teorie měření jsou posuzovací škály založeny na uspořádání nemetrických veličin, znakové a kategoriální systémy na sčítání.

Obecné problémy reliability

Každý, kdo se chce hlouběji poučit o reliabilitě pozorování, se setká s několika problémy. Záhy zjistí, že literatura o těchto otázkách je obtížně dostupná a omezuje se převážně na časopisecké články. Zkoumání reliability pozorování nemá v pedagogice dlouhou tradici; údaje o spolehlivosti výsledků získaných pozorováním se u nás nevyžadují, ačkoliv u didaktických a psychodiagnostických testů to bývá samozřejmé.

Není divu, vždyť sám pojem reliabilita se v pedagogice objevoval téměř jen v souvislosti s didaktickými testy. Jim byla přizpůsobena terminologie, statistické úvahy, vzorce. O tom, že je třeba studovat také reliabilitu pozorování, se v pedagogice začalo mluvit kolem roku 1963.[11]) Intenzívní výzkum reliability ve vlastním slova smyslu se rozběhl až po roce 1972.[12]) Do dnešních dnů bylo vyvinuto již několik desítek statistických postupů, každoročně přibývají další, modifikují se dřívější.

Volbu adekvátního postupu ztěžuje již sám počet vzorců. Další obtíže[13]) způsobuje skutečnost, že různí autoři používají pro tytéž proměnné odlišné označení, což uživatelům znesnadňuje orientaci v literatuře i srovnávání různých přístupů. Neodborník nepozná, zda a v čem se postupy liší. Navíc, i když jsou vzorce uváděny v plném znění, v textu často chybí zmínka o podmínkách jejich použití a způsobu interpretování spočítaných hodnot.

Vlastní výklad o měření reliability tedy začneme obecnějším pohledem na reliabilitu pozorování; teprve potom přejdeme k úvahám speciálním.

Na čem vlastně reliabilita pozorování pedagogického procesu závisí? Stručně řečeno především na třech činitelích:

a) Na samotném pozorovateli. Ukázalo se, že záleží jak na počtu pozorovatelů, tak na jejich vlastnostech a dovednostech. Výzkumy prokázaly, že spolehlivost pozorování bývá ovlivněna: apriorní zaujatostí či zkresleným očekáváním, s nímž pozorovatelé přistupují k pozorování; mírou jejich trénovanosti, zkušenosti s pozorováním, tendencí interpretovat si pokyny pro pozorování po svém, citlivostí ke zpětnovazebním informacím o kvalitě jejich pozorování, konzistentností a stabilitou hodnocení v čase, způsobem interpretace výsledků pozorování, odolností vůči únavě a nudě.[14])
b) Na pozorovací technice. Záleží na účelu nástroje (příprava budoucích učitelů, použití v běžné školní praxi, výzkum), na typu použitého nástroje (posuzovací škála, znakový systém, kategoriální systém), jednotce pozorování (pravidelné časové intervaly, změna činnosti, změna mluvčího, braní časových vzorků, změna tématu apod.), objektu pozorování (učitel, žák, žáci, učivo atd.), počtu a obsahu položek nástroje, správnosti a úplnosti pokynů aj.[15])
c) Na okolnostech pozorování. Záleží na vzorku škol, tříd, vyučovacích hodin; vzorku učitelů, žáků jako jednotlivců i kolektivu tříd, vyučovacím předmětu, sledovaném učivu; záleží na místě, počtu, délce pozorování, časovém zařazení v rámci školního roku, týdne, dne; na organizačních formách vyučování i sledovaných vyučovacích metodách.

Východiskem k měření reliability je následující úvaha. Výsledky, které získáme pozorováním pedagogické skutečnosti, obsahují vedle „správné", tj. „pravé" složky, i složku „chybovou". Výzkumný pracovník by měl usilovat o maximalizování „pravé" složky a o potlačení „chybové" složky na minimum. Převedeno do jazyka teorie měření: 1. *celkový získaný výsledek (X) se skládá z pravého skóru (T) a chybového skóru (E). Formálně zapsáno:* $X = T + E$ Tento první předpoklad doplňuje klasická teorie měření pedagogicko-psychologických jevů dalšími čtyřmi.[16]) Podrobněji nám definují základní vlastnosti pravého a chybového skóru:

2. *Střední hodnota chybových skórů je nula. Platí tedy:* $\varepsilon E = 0$.
3. *Korelace pravého a chybového skóru je nula. Platí:* $\rho (T,E) = 0$.
4. Korelace chybových skorů různých měření je nula. Platí: $\rho (E_1, E_2) = 0$.
5. Korelace chybového skórů jednoho měření a pravého skóru druhého měření je nula. Platí tedy: $\rho (E_1, T_2) = 0$.

Z výše uvedených pěti předpokladů vychází teorie měření, když se snaží zodpovědět otázku, nakolik jsou výsledky získané pozorováním přesné, spolehlivé, stálé, konzistentní, tj. reliabilní. Ke stanovení reliability pozorování se používá tzv. koeficientu reliability (ϱ_{XX}), který můžeme definovat třemi způsoby.

Buď jako čtverec korelace celkového a pravého skóru. Formálně zapsáno:

$$\varrho_{XX} = \varrho^2_{XT}$$

nebo jako podíl pravého a celkového rozptylu:

$$\varrho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

anebo jako korelaci dvou paralelních měření, resp. dvakrát opakovaného měření bez průvodních účinků na pozorovanou skutečnost.

$$\varrho_{XX} = \varrho_{XX'}$$

Důležité je, že všechny tři definice reliability jsou navzájem ekvivalentní. V dalším textu budeme používat především druhé a třetí definice.

Po tomto obecném úvodu můžeme přejít ke konkrétnímu výkladu. Řekli jsme, že reliabilita pozorování závisí mj. na okolnostech pozorování. Zmíněné okolnosti se však neberou vždy v úvahu ani při koncipování výzkumných projektů, ani při plánování hospitací a vyhodnocování získaných výsledků.

Vezměme jen otázku, která je klíčová pro organizování hospitací, inspekcí i výzkumů na školách. Kolikrát bychom měli pozorovat jednoho učitele (skupinu učitelů), abychom se na získané výsledky mohli spolehnout?

Z předběžných závěrů B. Rosenshina vyplývá, že pokud chceme porovnat dvě skupiny učitelů, stačí u každého učitele 1 — 2 pozorování a získáme poměrně spolehlivé průměrné ukazatele o každé skupině. Chceme-li získat spolehlivá data o učiteli jako jednotlivci, potom 1 — 2 pozorování stačí pouze na to, abychom poznali, jak učitel zvládá konkrétní (často jedinečnou) situaci. Pokud se zajímáme o kognitivní a snad i afektivní souvislosti učitelovy činnosti, bylo by třeba 10 — 30 pozorování.[17]) Z ryze statistického hlediska se k tomuto problému později vrátil G. Rowley[18]) a navrhl tři vzorce pro výpočet reliability pozorování (v závislosti na počtu, délce, počtu i délce pozorování současně). Jeho přístup by se dal stručně popsat takto. Během N pozorovacích period získáme postupně výsledky $X_1, X_2 \ldots X_i, X_j \ldots X_n$. Předpokládejme dvě věci:
1.Každé měření je standardizováno na jednotku rozptylu.
2.Korelace mezi dvěma výsledky měření X_i, X_j získanými jak v rámci jednoho

souvislého pozorování, tak v rámci několika nezávislých pozorování se nijak výrazně nemění.

Můžeme tedy uvažovat, že v rámci jednoho typu pozorování (např. souvislého, resp. několika nezávislých) jsou si korelace rovny. Vzorec pro výpočet reliability má potom tvar:

$$\varrho_{n,N} = \frac{nNr}{1 + (n-1)R + n(N-1)r}$$

kde $\varrho_{n,N}$ je hledaná reliabilita závislá na délce i počtu pozorování, n je délka pozorovací periody (délka časového úseku v minutách), N je počet pozorovacích period (po sobě jdoucích nezávislých úseků pozorování), r je reliabilita výsledků získaných při určitém počtu nezávislých pozorování. R je empiricky zjištěná korelace mezi výsledky získanými v rámci jednoho souvislého pozorování, přičemž platí $R \geq r$. Jestliže $N = 1$, potom se daná rovnice zjednoduší na rovnici pro závislost reliability na délce pozorování, pro $n = 1$ se zjednoduší na závislost reliability na počtu pozorování.

Reliabilita škálovacích technik

Teorie měření obvykle považuje škálování za pouhý předstupeň k měření v užším slova smyslu. Některé ze škálovacích postupů se tedy blíží pouhému kategorizování dat, zatímto jiné spíše vlastnímu měření.

O škálách lze uvažovat z mnoha hledisek. Nás budou především zajímat škály jako obecné metody, jichž lze použít pro kvantifikaci výsledků pozorování, a to na různých úrovních. Protože se označení „škála" používá i pro konkrétní metody (viz např. posuzovací škály, škály s kumulovanými body, škály typu nucené volby), užijeme pro obecné metody názvu kvantifikační stupnice, resp. stupnice. V literatuře se nejčastěji rozlišují čtyři druhy stupnic: nominální, ordinální, intervalové a poměrové. Podrobnější výklad o každé z nich nalezne čtenář např. v práci V. Břicháčka.[19])

Soustředíme se na stupnice ordinální, intervalové a nominální, neboť se jich při pozorování pedagogického procesu užívá nejčastěji.

Podívejme se nejprve na ordinální stupnice, jež umožňují porovnat každé dvě hodnoty posuzované vlastnosti objektu z hlediska pořadí. Odstupňování posuzované vlastnosti může být provedeno slovně, graficky, číselně atd. Je třeba připomenout, že u numerických ordinálních stupnic udávají jednotlivé číslice pouze umístění posuzované vlastnosti v daném pořadí, nikoli velikost rozdílu. U ordinálních stupnic se nepředpokládá, že by škála musela být „pravidelná"; vzdálenosti mezi dvojicemi sousedních bodů nemusejí být stejně veliké.

Někdy můžeme mít na škále dány i pevné body. Obvykle jsou to oba krajní póly (jde o tzv. zakotvení škály) a neutrální střed. O dvou výsledcích pozorování vyjádřených škálovou hodnotou můžeme uvažovat jen z hlediska pořadí (větší — menší než ..., častější — méně častý než ..., intenzívnější — méně intenzívní než ..., stejný jako ...). Nemůžeme však určovat, kolikrát je posuzovaná vlastnost větší, častější, intenzívnější.

Příklad
Několik pozorovatelů sleduje průběh výuky a pomocí sedmistupňové škály hodnotí, zda chování učitele vůči žákům bylo tvrdé (hodnota 1), či laskavé (hodnota 7), ani tvrdé, ani laskavé (hodnota 4). U učitele A vyjde průměrná hodnota škálové dimenze 2,1 a u učitele B 6,3. Můžeme říci, že učitel A je vůči žákům tvrdší než učitel B nebo že učitel B je vůči žákům laskavější než učitel A. Nemůžeme však prohlásit (a věcně by to ani nemělo smysl), že učitel A je třikrát přísnější než učitel B, resp. učitel B je třikrát laskavější než učitel A.

Druhým typem jsou tzv. intervalové stupnice. Umožňují porovnat každé dvě hodnoty posuzované vlastnosti nejen z hlediska pořadí, ale také stanovit velikost jejich rozdílu. To proto, že jde o „pravidelnou" škálu, vzdálenosti mezi dvojicemi sousedních bodů — intervaly — musí být stejně veliké. Někdy dokonce můžeme u těchto stupnic zavést i přirozený počátek — nulu a k ní potom vztáhnout všechny ostatní hodnoty. O dvou výsledcích pozorování vyjádřených škálovou hodnotou lze potom tvrdit, že hodnota *a* je dvakrát větší než hodnota *b* nebo že rozdíl mezi *a* a *b* je x-krát menší než mezi hodnotami *c* a *d*.

Podívejme se nyní, kterými statistickými postupy se u jednotlivých druhů stupnic má počítat spolehlivost zjištěných výsledků. Protože většina konkrétních posuzovacích škál užívaných při pozorování pedagogického procesu jsou stupnice ordinální, měli bychom pro výpočet reliability užít jen neparametrické statistiky. Obvykle se doporučuje počítat Kendallův koeficient shody W[20]), jenž může nabývat hodnot od 1,00 (při naprosté shodě posuzovatelů) do 0,00 (při naprosté neshodě). Protože koeficient shody W vyjadřuje vlastně průměrnou shodu mezi všemi posuzovateli, byly vyvinuty ještě další koeficienty. Uveďme např. Taylorův koeficient shody[21]), jenž z Kendallova vychází. Není zcela nový — specialisté upozorňují, že je jistou modifikací Spearmanova koeficientu pořadové korelace.

Podle J. B. Taylora můžeme průměrnou shodu mezi dvěma posuzovateli spočítat podle vzorce:

$$r_W = \frac{kW - 1}{k - 1}$$

kde k je počet posuzovatelů,
W je zjištěná hodnota Kendallova koeficientu shody.

Míru shody mezi dvěma skupinami posuzovatelů vypočteme podle vzorce:

$$r_{kW} = \frac{kW - 1}{(k - 1)W}$$

kde k a W značí totéž co v předchozím případě.

Testovacím kritériem Taylorova koeficientu shody je veličina chí-kvadrát, definovaná takto:

$\chi^2 = Wk(k - 1)$ při $(n-1)$ stupních volnosti

Uvedeného testovacího kritéria můžeme použít jen v případě, že $n > 7$ (kde n je počet posuzovaných objektů, vlastností atd.).

Řada odborníků se domnívá, že mnoho ordinálních stupnic používaných v pedagogice a psychologii se dost přibližuje stupnicím intervalovým. Abychom však mohli při výpočtech zacházet s ordinálními stupnicemi jako s inter-

valovými, musíme předpokládat, že intervaly u konkrétní používané škály jsou téměř stejné a rozdělení získaných dat přibližně normální. Výzkumy různých autorů ukazují, že oba předpoklady bývají často splněny, a proto lze v těchto případech použít při výpočtu reliability vydatnějších postupů parametrické statistiky.[22]) Zároveň je třeba upozornit, že při volbě statistických metod nelze tento závěr aplikovat mechanicky, ale splnění obou předpokladů bychom si měli ověřovat.[23])

Při měření reliability intervalových stupnic se dost často používá koeficientu vnitroskupinové, vnitrotřídní korelace (intrakorelace), který vychází z jednofaktorové analýzy rozptylu. Za zdroj „správného" rozptylu jsou brány jen posuzované objekty, vlastnosti. Novější experimentální výzkumy však prokazují, že věcně i statisticky budou výhodnější postupy opírající se o dvoufaktorovou analýzu rozptylu. Podrobnější popis doporučovaných modelů je uveden v poznámkách. [24])

Probrali jsme otázky reliability posuzovacích škál a zbývají nám znakové a kategoriální systémy.

RELIABILITA ZNAKOVÝCH A KATEGORIÁLNÍCH SYSTÉMŮ

Znakové a kategoriální systémy můžeme z hlediska teorie měření zařadit mezi nejjednodušší, tzv. nominální stupnice. Sledované znaky, vlastnosti objektu se v nich sice zapisují pomocí grafických symbolů, číslic, ale ty mají pouze funkci označovací. Např. zápis pomocí Flandersova formálního jazyka „...4, 9, 7, ..." nám jen říká, že se postupně objevily tři činnosti: učitelova otázka (4), žákova spontánní odpověď (9) a učitelův nesouhlas (7). Použité číslice nemají význam čísel, nemá smysl je sčítat, nemá smysl mluvit o tom, že jedna je větší než druhá, protože bychom říkali, že „jedna činnost je větší než druhá". Jde jen o „zkrácená jména" určitých vlastností a stavů objektu.

Pomocí znakových a kategoriálních systémů můžeme pouze zjišťovat, jak často se zmíněné stavy a vlastnosti vyskytly, případně jak po sobě následovaly. Zdálo by se, že pozorovatel, který sleduje, hodnotí a zapisuje určité pedagogické situace, má zde málo příležitostí dopouštět se chyb. Přesto však může dojít např. k nesprávnému určení druhu činnosti, vynechání některé činnosti, nesprávnému určení začátku a konce, nemluvě o situacích, které má pozorovatel nejen popisovat, ale i průběžně hodnotit. Jestliže tutéž pedagogickou situaci pozorují a zaznamenávají dva i více pozorovatelů, počet chyb obvykle vzrůstá. Je tedy důležité zjistit, zda výsledky pozorování nejsou chybami zatíženy natolik, že už vůbec nejsou spolehlivé.

Jakých statistických postupů můžeme použít při výpočtu reliability? Dříve než začneme odpovídat na tuto otázku, je nutné upozornit, že pojetí reliability nominálních stupnic se historicky vyvíjelo, a proto existují nejméně dvě odlišné odpovědi.

Zpočátku byl pojem reliabilita chápán pouze jako míra shody mezi dvěma a více pozorovateli. Čím více se pozorovatelé ve svých hodnoceních shodovali, tím určitěji se výsledkům připisovala přesnost a spolehlivost. Stačí však připomenout, že uvedené pojetí vlastně bere v úvahu jen pozorovatele, ačkoliv reliabilita závisí alespoň na třech činitelích (pozorovatelé, pozorovací techni-

ka, okolnosti pozorování). Proto se od sedmdesátých let začíná postupně prosazovat komplexnější pojetí, ze statistického hlediska složitější.

Uvedená dvě pojetí není nutné ostře stavět proti sobě. Jde o dvě vývojové etapy, z nichž každá má své výhody i nevýhody. Tam, kde řešíme jednodušší úlohy rutinní povahy, kde nezáleží tolik na komplexním přístupu, kde nemáme k dispozici vyšší výpočetní techniku a potřebujeme zjistit spolehlivost výsledků v krátké době, použijeme zřejmě jednodušších statistických přístupů. Patří k nim měření shody mezi pozorovateli, popřípadě vnitroskupinové korelace (intrakorelace). U výzkumných studií bychom měli použít složitějších statistických postupů, k nimž patří např. Cronbachova teorie zobecnitelnosti.

Měření shody mezi pozorovateli má ve statistice dlouhou tradici, počínaje pracemi G. U. Yula z let 1900 — 1912 a konče osmdesátými léty. Za oněch 70 — 80 let se v pedagogice a psychologii zavedlo a vyzkoušelo několik desítek různých koeficientů. [25]) S obdobnou situací se setkáváme i v dalších společenskovědních oborech, označovaných na Západě jako vědy o chování. [26]) O použití jednotlivých koeficientů nerozhodují jen jejich ryze statistické vlastnosti, ale také charakteristiky úloh, které potřebujeme prostudovat, dostupnost různých prostředků výpočetní techniky atd.

Podívejme se na podstatu problému, který máme řešit. Dejme tomu, že máme k dispozici kategoriální systém, jenž popisuje dění ve třídě jen třemi kategoriemi (mluví učitel, mluví žák, ticho nebo zmatek). Dva pozorovatelé kódují pomocí uvedeného systému tutéž vyučovací hodinu současně. Výsledky kódování lze potom psát do tabulky, která zachycuje míru jejich vzájemné shody (viz tab. 1).

Tab. 1 Kontingenční tabulka $C \times C$

		2. pozorovatel: kategorie C_1	C_2	C_3	marginální četnosti
1. pozorovatel (nebo vzor, kritérium): kategorie	C_1	n_{11}	n_{12}		n_{1+}
	C_2		n_{22}		n_{2+}
	C_3			n_{33}	n_{3+}
marginální četnosti		n_{+1}	n_{+2}	n_{+3}	N

n_{i+}

n_{+i}

Při kódování může nastat vždy jedna ze dvou možností. Buď oba kódují tutéž událost shodně, nebo rozdílně. První případ: 1. pozorovatel kóduje C_3 a 2. pozorovatel rovněž C_3 Shodli se, údaj bude ležet v políčku n_{33}. Druhý případ: 1. pozorovatel kóduje C_1, 2. pozorovatel však myslí, že už se vyskytla kategorie C_2 Neshodli se, údaj bude ležet v políčku n_{12} Čím více údajů leží na hlavní diagonále matice (tj. v políčku n_{11}, n_{22} n_{33}, obecně n_{ii}), tím větší je shoda mezi pozorovateli. Marginální (okrajové) četnost nám pak udávají součet údajů v každém řádku matice (obecně n_{i+}) nebo v každém sloupci matice (n_{+i}).

Nyní potřebujeme zjistit míru vzájemné shody obou pozorovatelů. Který statistický postup máme zvolit? Začneme negativním vymezením. Neměli bychom používat chí-kvadrát testu, neboť je pro tyto účely nevhodný. I když se někdy mluví o chí-kvadrát testu dobré shody, musíme si uvědomit, o jakou shodu jde.

Zmíněný test totiž předpokládá, že rozdělení dat v kontingenční tabulce je „rovnoměrné", je ve statistickém smyslu náhodné a statisticky nezávislé. Chí-kvadrát test zjišťuje, zda se empiricky získaná data shodují s tímto modelovým předpokladem či nikoli. Jedna z podmínek praktického použití zmíněného testu říká, že žádná teoreticky očekávaná četnost nesmí být menší než 1 a v tabulce nemá být více než 20% očekávaných četností menších než 5. Konečně je třeba připomenout, že míry statistické závislosti odvozené od chí-kvadrát testu nejsou jednoznačně definovány; jejich návaznost na chí-kvadrát test je dost libovolná. [27])

My však předpokládáme jistou zákonitost v rozdělení dat, předpokládáme určitou shodu mezi pozorovateli, kteří pozorovali tentýž pedagogický proces, a chceme zjistit míru jejich vzájemné shody.

V zásadě jsou možné dva přístupy. První vychází ze skutečně zjištěné shody a neshody mezi pozorovateli:

$$\% \text{ shody} = \frac{\text{skutečná shoda}}{\text{skutečná shoda} + \text{skutečná neshoda}} \times 100$$

Druhý, přesnější přístup vychází ze vztahu mezi skutečně zjištěnou a teoreticky očekávanou shodou:

$$\text{koeficient} = \frac{\text{skutečná proporce shody} - \text{očekávaná proporce shody}}{1 - \text{očekávaná proporce shody}}$$

Z mnoha různých koeficientů je ovšem třeba vybrat ten, který je adekvátní dané situaci. Rozhodování o tom, kdy máme kterého z nich použít, nám usnadní přehledná tabulka T. Fricka a M. I. Semmela. [28]) Viz tab. 2. Vybrané koeficienty shody si teď rozebereme podrobněji. [29])

Jedním z nejjednodušších je koeficient marginální shody. Zjišťujeme jím přesnost pozorování nějakého pozorovatele vzhledem ke kritériu, ke standardu. Čím více se pozorovatel blíží standardu (údajům vzorového pozorování), tím více se hodnota koeficientu blíží jedné. Čím více se odchyluje od standardu, tím více se hodnota koeficientu blíží nule.

Tab. 2
Tabulka pro výběr vhodného statistického postupu (podle T. Fricka a M. I. Semmela)

	Shoda mezi pozorovateli			Reliabilita
	kriteriální	intraindivid.	interindivid.	
předpoklad	přesnost kódování vzhledem ke kritériu (normě)	shoda pozorovatele se sebou samým; konzistentnost	shoda mezi dvěma a více pozorovateli	míra rozlišitelnosti mezi subjekty vzhledem k rozdílům uvnitř subjektů a dalším chybám
kdy použít	před a v průběhu shromažďování dat; výcvik pozorovatele	před a v průběhu shromažďování dat	po shromáždění dat	po shromáždění dat
prostředek	videozáznam (vybrané jednoznačné ukázky)	videozáznam (situace reprezentativní pro běžné kódování)	živé pozorování ve třídě	živé pozorování ve třídě
statistika pro kategoriální systémy	Lightův koef. kappa; koef. marginál. shody	Flandersova modifikace koeficientu π	vnitroskupinová korelace (intrakorelace)	Cronbachova teorie zobecnitelnosti
statistika pro znakové systémy	Cohenův koeficient kappa	Cohenův koeficient kappa	vnitroskupinová korelace (intrakorelace)	Cronbachova teorie zobecnitelnosti

Může nás zajímat, jak dalece se pozorovatel shoduje se standardem při pozorování jedné kategorie činnosti (obecně i-té kategorie).
Příslušný vzorec má tvar:

$$\bar{P}_{oi} = \frac{\min(n_{i+}, n_{+i})}{\max(n_{i+}, n_{+i})},$$

kde n_{+i}, n_{i+} jsou marginální četnosti pro i-tou kategorii, zlomek min/max značí, že nejmenší hodnotu marginální četnosti dělíme největší hodnotou marginální četnosti.
Doporučovaná hodnota koeficientu $\bar{P}_{oi}$ je 0,90 nebo vyšší.

Obvykle nás ovšem zajímá celková shoda pozorovatele se standardem, míra shody ne u jedné, nýbrž u všech sledovaných kategorií činnosti. Vypočítáme ji jako průměr všech dílčích hodnot $\bar{P}_{oi}$:

$$\bar{P}_o = \frac{1}{C} \sum_{i=1}^{C} \bar{P}_{oi}$$

kde C je počet kategorií pozorovacího systému,
$\bar{P}_{oi}$ jsou hodnoty vypočtené podle předchozího vzorce.

Poněkud složitější je další koeficient shody mezi pozorovateli. Je jím Flandersova modifikace Scottova koeficientu π. Flandersova koeficientu π_f používáme tehdy, když nás zajímá, jak konstantní a stabilní je pozorování téhož pozorovatele (zda tytéž situace kóduje s odstupem času shodně nebo rozdílně). Hodnota koeficientu p_f se opět pohybuje od nuly do jedné; nula značí maximální neshodu, jedna pak dokonalou shodu. Příslušný vzorec má tvar:

$$\pi_f = \frac{P_{of} - P_{ef}}{1 - P_{ef}} \qquad \text{kde} \qquad P_{of} = 1 - \sum_{i=1}^{C} \left| \frac{n_{+i}}{N} - \frac{n_{i+}}{N'} \right|$$

$$P_{ef} = \frac{1}{4} \sum_{i=1}^{C} \left(\frac{n_{+i}}{N} + \frac{n_{i+}}{N'} \right)^2$$

přičemž C je počet kategorií pozorovacího systému $(C \geq 2)$
n_{+i}, n_{i+} marginální četnosti pro i-tou kategorii
N, N' je celkový počet činností okódovaných oběma pozorovateli (součet marginálních četností sloupců nebo řádků — viz tab. 1).
Rozlišení mezi N a N' je nutné tehdy, když se celkový počet zaznamenaných kategorií činnosti u obou pozorovatelů liší. Jde o případy, kdy jeden z pozorovatelů zapomene kódovat, vynechá některou činnost a celkově zaznamená méně činností než druhý.

Doporučená hodnota Flandersova koeficientu π_f by měla být vyšší než 0, 70, popřípadě 0, 75.

Třetí koeficient shody mezi pozorovateli je Cohenův koeficient kappa. Používáme ho u znakových systémů, pokud nás zajímá shoda pozorovatele s kritériem, standardem (doporučená hodnota je 0,80 a vyšší) anebo konzistentnost, stabilita pozorování u téhož pozorovatele (doporučená hodnota 0,75 a vyšší). Hodnota koeficientu kappa se pohybuje od -1 do 1 (od úplné neshody po naprostou shodu). Vzorec je vlastně jinou modifikací Scottova koeficientu π a má tvar:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

kde $P_o = \frac{1}{N} \sum_{i=1}^{C} n_{ii}$ (zjištěná proporce shody)

$$P_e = \frac{1}{N^2} \sum_{i=1}^{C} (n_{i+})(n_{+i}) \text{ (očekávaná proporce shody)}$$

přičemž N, C, n_{+i}, n_{i+} mají stejný význam jako v předchozích případech a n_{ii} je četnost prvku ležícího na hlavní diagonále kontingenční tabulky.

Jako poslední koeficient shody uvedeme Lightovu modifikaci právě zmíněného Cohenova kappa. Lightova koeficientu kappa užíváme u kategoriálních systémů, pokud nás zajímá shoda pozorovatele s kritériem, standardem. Může však mít i bohatší použití, neboť ho po úpravě lze použít k měření shody více pozorovatelů se standardem, k měření podmíněné shody mezi dvěma a více pozorovateli, k zjišťování shodných posloupností dat v pozorování několika pozorovatelů apod. Výchozí Lightův vzorec má tvar:

$$\kappa = 1 - \left(\frac{d_o}{d_e}\right)$$

kde $d_o = 1 - P_o$ (zjištěná proporce neshody)
$d_e = 1 - P_e$ (očekávaná proporce neshody)
přičemž P_o, P_e označují stejné výrazy jako u předchozího vzorce kappa.

Výhodné je, že u Cohenova i Lightova koeficientu lze testovat statistickou významnost zjištěných hodnot. Vypočtené hodnoty však musíme transformovat tak, aby se empirické rozdělení blížilo normálnímu. Příslušnou hodnotu Z vypočteme takto:

$$Z = \frac{\kappa - \kappa_0}{\sqrt{\text{var}(\kappa)}},$$

kde κ je vypočtená hodnota koeficientu κ,

$$\text{var}(\kappa) = \frac{P_0(1-P_0)}{N(1-P_e)^2}$$

κ_0 je hodnota koeficientu κ očekávaná v souladu s výzkumnou hypotézou. Pokud je hodnota κ_0 podle nulové hypotézy rovna nule (tj. předpokládáme, že shoda mezi pozorovateli je pouze náhodná), vzorec pro Z se zjednoduší na tvar:

$$Z = \frac{\kappa}{\sqrt{\frac{P_e}{N(1-P_0)}}}$$

Všechny užité symboly mají shodný význam jako dříve. Je však důležité připomenout, že rozdělení se blíží normálnímu jen při větším počtu pozorovacích dat.

Probrali jsme některé důležité koeficienty, jimiž lze měřit shodu mezi pozorovateli. Nakonec nám zbývají složitější postupy pro zjišťování reliability.

V oddíle o obecných otázkách reliability jsme uvedli, že koeficient reliability lze definovat třemi různými způsoby, mj. jako podíl pravého, „správného" rozptylu a celkového rozptylu. Tedy:

$$\varrho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Tato definice sice naznačuje, jak asi dospět k výpočtu reliability, ale neříká nic o tom, co máme považovat za zdroj „chybového" rozptylu.

Už v roce 1963 uvažovali D. Medley a H. Mitzel o možných zdrojích chyb při systematickém pozorování pedagogického procesu. Vzali v úvahu nejen rozdíly mezi pozorovateli (jak je studují výše diskutované koeficienty shody mezi pozorovateli, ale ještě další rozdíly: mezi učiteli, mezi situacemi, v nichž byla činnost učitele a žáků sledována, a mezi pozorovacími technikami (zajímali se o počet sledovaných položek). Jako možné zdroje „chybového" rozptylu tedy zvolili čtyři faktory a pro výpočet koeficientu reliability navrhli a vyzkoušeli čtyřfaktorovou analýzu rozptylu. [30])

Jejich přístup byl v oné době nesporným pokrokem, uvažoval o spolehlivosti pozorování komplexněji než tradiční koeficienty shody mezi pozorovateli. Za čas se však zjistilo, že se opírá o statistický předpoklad, který z pedagogicko-psychologického hlediska není zcela správný. Jmenovaní autoři totiž předpokládali, že nestabilita chování učitele a žáků v různých situacích je způsobena pouze náhodnými chybami, pramenícími buď v lidech, nebo v prostředí. Předpokladali, že měřené charakteristiky jsou v principu stálé a nepodléhají zákonitým změnám.

Rozpor mezi skutečností a statistickým předpokladem bylo ovšem třeba

odhalit a vyřešit. Museli jsme si počkat až do r. 1972, kdy Mc Gaw se spolupracovníky nejen upozornili, že zmíněný předpoklad neplatí, ale navrhli vhodnější statistický postup. [31])

Tímto postupem je statistická teorie L. J. Cronbacha a spolupracovníků [32]) z let 1963 — 1972. V angličtině se nazývá „theory of generalizability"; J. Kalous navrhl český ekvivalent „teorie zobecnitelnosti" [33]). Výborný přehled celé teorie, metodologických problémů s ní spojených i dosavadních praktických aplikací za léta 1973 — 1980 podávají R. J. Shavelson a N. M. Webbová. [34]) I když teorie zobecnitelnosti našla použití v mnoha oblastech pedagogiky a psychologie [35]), autoři uvedeného přehledu spatřují její hlavní přínos v tom, že umožňuje komplexněji změřit spolehlivost výsledků pozorování.

Ze statistického hlediska vychází Cronbachova teorie zobecnitelnosti z analýzy rozptylu, ale v některých směrech se od ní liší. Předpokládá, že každé pozorování je jen určitým vzorkem z celého univerza možných pozorování, a tvrdí, že musíme vzít v úvahu komplexnost problému, máme-li výsledky správně zobecnit.

Univerzum možných pozorování závisí na kombinaci různých podmínek, za nichž se pozorování pedagogického procesu uskutečňuje. (V terminologii teorie zobecnitelnosti se tyto podmínky označují jako „fasety".) Podle charakteru problému si výzkumný pracovník vybere rozhodující podmínky a tím si vymezí univerzum přípustných podmínek, za nichž se budou výsledky získávat, vyhodnocovat a zobecňovat.

Získané výsledky, tj. celkový rozptyl dat, budou opět obsahovat „správnou" složku a složku „chybovou". „Správná" složka však nyní nebude totožná s pravým skórem jako u tradiční teorie měření, nýbrž bude střední hodnotou skórů ze všech měření přípustných pro danou kombinaci podmínek, pro definované univerzum. (V terminologii teorie zobecnitelnosti se tato „správná" složka rozptylu nazývá „univerzální skór".)

Teorie zobecnitelnosti podstatně přispěla i k rozpracování „chybové" složky. Předpokládá totiž, že každá kombinace podmínek s sebou přináší i kombinaci možných chyb měření, a tvrdí, že chyby vznikající při měření mají multidimenzionální povahu. Proto i „chybovou" složku rozptylu je třeba chápat multidimenzionálně a je nutno ji rozložit na jednotlivé části.

Cronbachova teorie zobecnitelnosti chápe otázku reliability spíše jako přesnost zobecnění. Umožňuje vypočítat tzv. koeficient zobecnitelnosti, jenž je definován jako podíl mezi rozptylem univerzálního skóru a empiricky zjištěným skutečným rozptylem. Koeficient zobecnitelnosti je obsahově srovnatelný s tradičními koeficienty reliability (zejména s koeficientem vnitrotřídní korelace), výpočtově je ovšem značně náročnější.

Dosavadní zkušenosti naznačují, že při rutinním pozorování v pedagogické praxi zatím Cronbachova teorie nenašla širší použití — dává se přednost jednodušším statistickým postupům. Přibývají však výzkumné práce, které využívají teorie zobecnitelnosti nejen ve výše uvedeném komplexním pojetí, ale také při řešení speciálních problémů: při měření konzistence a stability pozorování u jednoho pozorovatele, při zjišťování shody mezi pozorovateli [36]), při měření spolehlivosti pozorování u posuzovaných škál. [37])

ZÁVĚR

Jednou z důležitých metodologických otázek pedagogického pozorování je jeho přesnost, spolehlivost, stálost, konzistentnost. Přehledová studie se pokusila rozebrat jednu z důležitých statistických charakteristik standardizovaného pozorování, jeho reliabilitu.

Ukazuje se totiž, že reliabilita pozorování závisí jak na pozorovateli samotném, tak na zvolené pozorovací technice i na okolnostech pozorování. Z hlediska uživatele je důležité vědět, že neexistuje „nejlepší" nebo univerzální statistický postup, nýbrž že je třeba zvolit postup nejvhodnější pro daný případ. Proto byly podrobněji vyloženy podmínky použití různých statistických postupů pro výpočet reliability u posuzovacích, znakových i kategoriálních pozorovacích technik.

V článku jsme se snažili dodržet hlavní linii výkladu a ponechali jsme stranou některé speciální otázky. Např. zjišťování shodných posloupností dat u několika pozorovatelů, konkrétní postupy používané v teorii zobecnitelnosti apod.

Studie chce přispět k řešení dílčího metodologického problému, jemuž v socialistických státech nebyla zatím věnována větší pozornost, i když je důležitý pro pedagogickou teorii i praxi.

POZNÁMKY

[1]) Skalková, J. a kol.: *Stav a perspektivy rozvoje vědního oboru pedagogika.* (Podkladový materiál pro rozšířené zasedání vědeckého kolegia pedagogiky a psychologie ČSAV). Praha, PÚ JAK ČSAV 1980. Monoszon, E. I.: *K problematike issledovanij po metodologii pedagogiki.* Sovetskaja pedagogika 1981, č. 12, s. 39.

[2]) Vorobjev, G. V.: *Uslovija effektivnosti eksperimentu v pedagogike.* Sovetskaja pedagogika 1981, č. 5, s. 100.

[3]) Kerlinger, F. N.: *Základy výzkumu chování.* Praha, Academia 1972, s. 434.

[4]) Skatkin, M. N.: *Osnovnyje napravlenija issledovanija problemy izučenija, obobščenija i ispoľzovanija peredovogo pedagogičeskogo opyta.* Sovetskaja pedagogika 1979, 2, s. 20; Bitinas, B.: *Ispoľzovanije statističeskogo metoda v izučenii i obobščenii pedagogičeskogo opyta.* In: Monoszon, E. I. (Red.): *Metodologičeskie i teoretičeskie problemy izučenija, obobščenija i ispoľzovanija peredovogo pedagogičeskogo opyta.* Moskva, APN SSSR 1978, s. 62—63.

[5]) Osipov, G. V.: *Vstupiteľnaja staťja.* In: *Matematičeskije metody v sovremennoj buržoaznoj sociologii.* Moskva, Progress 1966, s. 5—20.

[6]) Kuzmina, N. V.: *Metody issledovanija pedagogičeskoj dejateľnosti.* Leningrad, LGU 1970, s. 65; Rys, S.: *Hospitace v pedagogické praxi.* Praha, SPN 1975, s. 48; Pergler, P. a kol.: *Vybrané techniky sociologického výzkumu.* Praha, Svoboda 1969, s. 180—182.

[7]) Zpravidla má podobu tištěného materiálu (znění techniky a instrukce pro uživatele), někdy doplněného o zácvikový program, instruktážní film, videozáznam vybraných situací apod.

[8]) Ukázka ze škály D. Ryanse. Viz Kerlinger, F. N.: op. cit. s. 653

[9]) Ukázka techniky B. Browna. Cit. podle: Borisch, G. D. — Fenton, K. S.: *The Appraisal of Teaching. Concepts and Process.* Reading, Addison-Wesley 1977, s. 17.

[10]) Modifikovaná ukázka techniky N. A. Flanderse. Cit. podle: Flanders N. A.: *Analyzing Teaching Behavior.* Reading, Addison-Wesley 1970, s. 34.

[11]) Medley, D. M. — Mitzel, H. E.: *Measuring Classroom Behavior by Systematic Observation.* In: Cage, N. L. (Ed.): *Handbook of Research on Teaching.* Chicago, Rand Mc Nally 1963, s. 309—328.

[12]) Mc Gaw B. — Wardrop, J. L. — Bunda, M. A.: *Classroom Observation Schemes: Where the Errors?* Amer. Educ. Research Journal, roč. 9, (1972), č. 1, s. 13—27.

[13]) Podrobněji je rozebírají např. House, A. E. — House, B. J. — Campbell, M. B.: *Measuring of Interobserver Agreement: Calculation Formulas and Distribution Effects.* Behavioral Assessment, roč. 3 (1981), č. 1, s. 37 n.

[14]) Wasik, B. H. — Loven, M. D.: *Classroom Observational Data: Sources of Inaccuracy and Proposed Solutions.* Behavioral Assessment roč. 2 (1980), č. 3, s. 217—220.

[15]) Viz podrobněji: Herbert, J.—Attridge, C.: *A Guide for Developers and Users of Observation Systems and Manuals.* Amer. Educ. Research Journal, roč. 12 (1975), č. 1, s. 1—20.

[16]) Cit. podle práce Kalous, J.: *Klíčové psychometrické problémy v současné teorii testů.* Čs. psychologie, roč. 22 (1978), č. 4, s. 351—352.

[17]) Rosenhine, B.: *The Smallest Meaningful Sample of Classroom Transaction.* Journal of Research in Science Teaching, roč. 10 (1973), č. 3, s. 221—226.

[18]) Rowley, G.: *The Relationship of Reliability in Classroom Research to the Amount of Observation: An Extension of the Spearman — Brown Formula.* Journal of Educational Measurement, roč. 15 (1978), č. 3, s. 165—180.

[19]) Břicháček, V.: *Úvod do psychologického škálování.* Bratislava, Psychodiagnostické a didaktické testy, n. p., 1978, s. 17—25.

[20]) Viz např. Kerlinger, F. N.: op. cit. s. 270 n.

[21]) Cit. podle Bintig, A.: *The Efficiency of Various Estimations of Reliability of Rating Scales.* Educational and Psychological Measurement, roč. 40 (1981), č. 3, s. 629—630.

[22]) Na základě svých experimentálních výzkumů dospěli k tomuto názoru např. S. Siegel, N. H. Anderson, F. H. Silverman. Převzato ze studie A. Bintiga, op. cit. s. 620.

[23]) Obezřetnost při volbě statistického postupu doporučují F. Kerlinger: op. cit. s. 417—420; V. Břicháček: op. cit. s. 43—44; aj.

[24]) Pro výpočet reliability doporučuje A. Bintig (op. cit. s. 622 — 624) dva následující modely. Rozdíl mezi nimi je v tom, že v prvním modelu jsou posuzovatelé chápáni jako náhodný výběr z jisté populace, zatímco v druhém modelu jde o konkrétní, stále stejné posuzovatele; náhodné jsou objekty, které mají být posuzovány. Odtud plyne i rozdílné užití. Prvního modelu použijeme při výzkumu toho, jak osoby vnímají určitou situaci. Druhý model je vhodnější pro klinické použití a případové studie. První statistický model předpokládá, že **n** individuálních objektů, vlastností (chápaných jako náhodné) jsou posuzovány **k** posuzovateli (uvažovanými jako náhodní) pomocí m-stupňové fixní posuzovací škály. Formálně zapsáno: $X_{ij} = \mu + a_i + \sigma_j + (ab)_{ij} + e_{ij}$
kde μ = populační průměr; průměr všech potenciálních pozorování; a_i = náhodná proměnná i-tého individuálního objektu, s normálním rozdělením $(0, \sigma_i^2)$

b_j = náhodná proměnná j-tého posuzovatele s normálním rozdělením $(0, \sigma_j^2)$

$(ab)_{ij}$ = interakce; náhodná proměnná posouzení *(ij)* s normálním rozdělením $(0, \sigma_{ij}^2)$

e_{ij} = posuzovací chyba vznikající při posouzení (ij) s normálním rozdělením $(0, \sigma_e^2)$ nezávislým na a_i, b_i

Celkový rozptyl X_{ij} můžeme zapsat jako:

$$\text{var}(X_{ij}) = \sigma_i^2 + \sigma_j^2 + \sigma_{ij}^2 + \sigma_e^2$$

Hlavním problémem je nyní odhadnout tu složku rozptylu, kterou nemůžeme změřit přímo. Abychom mohli spočítat reliabilitu, musíme rozptyl mezi individuálními objekty (σ_i^2) vztáhnout k celkovému rozptylu:

$$\varrho = \sigma_i^2 / (\sigma_i^2 + \sigma_j^2 + \sigma_{ij}^2 + \sigma_e^2)$$

Příslušná tabulka zdrojů rozptylu vypadá takto:

Zdroj rozptylu	*df*	*Průměr čtverců* MS	E(MS)
individuální objekty I	$n-1$	MSI	$\sigma_e^2 + k\cdot\sigma_i^2$
posuzovatelé J	$k-1$	MSJ	$\sigma_e^2 + \sigma_{ij}^2 + n\cdot\sigma_j^2$
interakce I × J	$(n-1)(k-1)$	MSIJ	$\sigma_e^2 + \sigma_{ij}^2$
chyba	0	(MSE)	σ_e^2

celkem $nk-1$

Dále platí

$$\sigma_i^2 = (\text{MSI} - \text{MSIJ})/k$$

$$\sigma_j^2 = (\text{MSJ} - \text{MSIJ})/n$$

$$\sigma_e^2 + \sigma_{ij}^2 = (\text{MSIJ})$$

Z toho už můžeme odvodit koeficient reliability pro jednoho posuzovatele:

$$r_{1,rand} = (\text{MSI} - \text{MSIJ}) / [\text{MSI} + (k-1)\,\text{MSIJ} + \frac{k}{n}(\text{MSJ} - \text{MSIJ}$$

Pro k posuzovatelů má koeficient reliability tvar:

$$r_{k,rand} = [n(\text{MSI} - \text{MSIJ})] / (n\,\text{MSI} + \text{MSJ} - \text{MSIJ}).$$

Druhý statistický model předpokládá, že **n** individuálních objektů, vlastností (chápaných jako náhodné) je posuzováno **k** fixními posuzovateli (jde vždy o tytéž posuzovatele) pomocí m-stupňové posuzovací škály.
Potom platí:

$$X_{ij} = \mu + a_i + \beta_9 + (a\beta)_{ij} + e_{ij}$$

kde je fixován model náhodných efektů β_j tak, že nedochází ke změně:

$$\text{var}\,(X_{ij}) = \sigma_i^2 + \sigma_{ij}^2 + \sigma_e^2$$

Abychom mohli spočítat reliabilitu, musíme rozptyl mezi individuálními objekty, vlastnostmi vztáhnout k celkovému rozptylu:

$$\varrho = \sigma_i^2 / \sigma_i^2 + \sigma_{ij}^2 + \sigma_e^2$$

Odhady, které by odpovídaly níže uvedené tabulce zdrojů rozptylu, jsou nyní obtížnější, neboť musíme samostatně vyčlenit složku MSE. Tuto chybu rozptylu však není snadné odhadnout. Protože odpovídající četnost je rovna jedné, stupeň volnosti se zde rovná 0.

Zdroj rozptylu	*df*	*Průměr čtverců* MS	E(MS)
individuální objekty I	$m-1$	MSI	$\sigma_e^2 + \sigma_{ij}^2 + k\cdot\sigma_i^2$
posuzovatelé J	$k-1$	MSJ	$\sigma_e^2 + \sigma_{ij}^2 + n\cdot\sigma_j^2$
interakce I × J	$(n-1)(k-1)$	MSIJ	$\sigma_e^2 + \sigma_{ij}^2$
chyba	0	(MSE)	σ_e^2

celkem $nk-1$

Koeficient reliability pro jednoho posuzovatele má potom tvar:

$$r_{1,fix} = (\text{MSI} - \text{MSIJ} + \sigma_{ij}^2) / [(\text{MSI} + (k-1)\,\text{MSIJ} + \sigma_{ij}^2].$$

Přitom je nulová jen mterakce rozptylů (σ_i^2) ; MSIJ je jistá chyba, jejíž velikost je volně odhadnuta. Za těchto podmínek můžeme koeficient reliability zjednodušit takto:

$$r'_{1,fix} = (\text{MSI} - \text{MSIJ}) / [\text{MSI} + (k-1)\,\text{MSIJ}]$$

Koeficient reliability pro **k** posuzovatelů má za uvedených podmínek tvar:

$$r'_{1,fix} = 1 - \text{MSIJ} / \text{MSI}$$

A. Bintig poznamenává, že ve zmíněné podobě je odhad reliability podle koeficientu r' poněkud nadsazen. Zvolíme-li jiný přístup (např. položíme-li $\sigma_{ij}^2 = \text{MSIJ}$), bude hodnota vypočteného koeficientu poněkud nižší, než by měla být. Za této situace doporučuje, abychom dali přednost koeficientu r'_1.

[25]) Jen výběrová studie A. E. House jich uvádí 20. Viz House, A. E. — Houce, B. J. — Campbell, M. B.: op. cit. s. 42—44.

[26]) V r. 1977 probíhala na stránkách časopisu Journal of Applied Behavioral Analysis dlouhá diskuse o měření shody mezi pozorovateli, aniž byla jednoznačně uzavřena.

Přehled dosavadních používaných koeficientů v ní podal Kelly. Kelly, M. B.: *A Review of the Observational Data Collection and Reliability Procedures Reported in JABA.* Journal of Applied Behavioral Analysis, roč. 10, 1977. s. 97—101.

[27]) Podrobnější výklad viz: Řehák, J. — Řeháková, B.: *Měření statistické závislosti nominálních znaků.* Sociologický časopis, roč. 9 (1972), č. 4, s. 404—418.

[28]) Citováno podle Frick, T. — Semmel, M. I.: *Observer Agreement and Reliabilities of Classroom Observational Measures.* Review of Educational Research, roč. 48, 1978, č. 1, s. 181.

[29]) Vzorce jsou uváděny podle práce T. Fricka a M. I. Semmela: op. cit. s. 167—172 (resp. podle starší Lightovy práce z r. 1971). Podrobnější statistické údaje jsou převzaty z rozsáhlé studie Light, R. J.: *Issues in the Analysis of Qualitative Data.* In: Travers, R. M. (Ed.) *Second Handbook Research on Teaching,* Chicago, Rand Mc Nally 1973, s. 331 an.

[30]) Viz poznámka 11.

[31]) Viz poznámku 12.

[32]) Cronbach, L. J. — Gleser, G. C. — Nanda, H. — Rajaratmam, N.: *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York, Wiley 1972.

[33]) Viz poznámku 16. op. cit. s. 353.

[34]) Shavelson, R. J. — Webb, R. J.: *Generalizability theory 1973 — 1980.* Brit. Journal of Mathematical and Statistical Psychology, roč. 34, 1981, č. 2, s. 133—166.

[35]) Např. u psychologických a didaktických testů, dotazníků apod. Kane, M. T.: *The Generalizability of Class Means. Review of Educational Research,* roč. 47, 1977, č. 1, s. 267—292.

[36]) Booth, C. L. — Mitchel, S. K. — Solin, F. K.: *The Generalizability Study as a Method of Assessing Intra- and Interobserver Reliability in Observational Research.* Behavioral Research Methods Instrumentation, roč. 11, 1979, č. 5, s. 491—494. Nešlo o ryze pedagogické situace, ale o interakci matky s kojencem.

[37]) Smith, P. L.: *The Generalizability of Student Ratings of Courses: Asking the Right Questions.* Journal of Educational Measurement, roč. 16, 1979, č. 2, s. 77—78.

ИРЖИ МАРЕШ

КАК УСТАНАВЛИВАТЬ НАДЕЖНОСТЬ НАБЛЮДЕИИЯ

Наблюдеине бывает важиым компонентом многнх педагогических исследоваиий, является базой пробиых уроков и ннспекцин. Однако в социалистических странах педагогика до настоящего времеии уделяла мало виимания степени, в какой результаты паблюдений страдают ошибками, мало времеии уделяла вопросу изыскаиия статистическнх методов устаиовлеиия надежности стандартнзированиого наблюдения.

В своем обзорном очерке, который опирается на статистические работы зарубежных авторов, указывает паправление решеиия этой иитересиой методологической проблемы. Сиачала дается определение основных понятий (стаидартнзированиое наблюдеиие, техника наблюдения, оценочная шкала, система призиаков деятельиости, система категорнй деятельности). Потом приведены три эквивалеитиые определения т. и. коэффициеита надежностн. Автор констатирует, что надежность наблюдеиия зависит от самого наблюдающего, от избраниой техники наблюдення и от обстоятельств наблюдеиия. Обсуждается вопрос: сколько раз следует наблюдать, чтобы мож-

ио было положиться иа полученные результаты?

Детальиому анализу подвержены статистические методы, пригодные для определения надежности оценочных шкал, систем прнзнаков н категорий деятельности (включительно условий примепения и рекомендуемых величии отдельных коэффициентов). У оценочных шкал рекомендуется пользоваться коэффициентом совпадения W (Kendall) и коэффициентом совпадения по Тейлору (при паличии порядковых шкал), или коэффициентом корреляции внутри групп (при паличии интервальных шкал). У систем призиаков и категорнй деятельности рекомендуется (в зависимости от разиых условий, указаиных в таб. № 2) применять: коэффицнент граничного совпадения, модификацию коэффициента (Flanders), коэффициент каппа (Cohen), модифицированный коэффициент каппа (Leight). В сжатой форме изложен также принцип теории обобщаемости (Cronbach) (theory of generalizability).

В заключение автор предупреждает пользователей, что для вычисления падежности наблюдения пе был еще разработан нн «наилучший», ни «универсальный» статистический метод, вследствие чего необходимо придерживаться порядка действий, который для данного конкретного случая представляется наиболее рациональным.

JIŘÍ MAREŠ

WAYS OF MEASURING RELIABILITY OF OBSERVATION

Observation is one of the important parts of many pedagogical researches, and is the basis of supervision and inspection. In the socialist countries, however, pedagogy has so far paid little attention to the extent of mistakes affecting the results of the observation, and to the question of what statistical methods should be used in verifying the reliability of standardized observation.

In a summary study, which is based on statistical works published abroad, the author suggests a way of solving this interesting methodological problem. First of all, definitions are given of the basic concepts (standardized observation, observational technique, rating scale, sign system, category system). Then three equivalent definitions are given of the coefficient of reliability. The reliability of observation is found to be dependent on the observer himself, on the observational technique chosen and on the circumstances of observation. How many times a teacher should be observed so that the results obtained can be relied upon is a question that is discussed here.

A detailed analysis is made of statistical methods suitable for measuring the reliability of rating scales, sign and category systems (including the conditions for use and the recommended values for each of the coefficients). What is recommended as regards rating scales is the use of Kendall's coefficient of agreement W and Taylor's coefficient of agreement (in the case of ordinal scales) or the intraclass correlation coefficient (in the case of interval scales). As regards sign and category systems, what is recommended (depending on various conditions specified in Table 2) is the use of: the coefficient of marginal agreement, Flanders's modification of coefficient π, Cohen's coefficient kappa, Light's modification of coefficient kappa. A brief explanation is also given of the principle of Cronbach's theory of generalizability.

In conclusion, users are reminded that there is no „best" nor „universal" statistical method to calculate the reliability of observation, but the method to be chosen is the one that is the most suitable for the case in question.