Kvalita dat II.*

Přístupy ohodnocování výzkumných instrumentů založené
na modelování kovariančních struktur

JAN ŘEHÁK**

Sociologický ústav AV ČR, Praha

Quality of Data II.

Approaches to the Evaluation of Research Instruments based on Covariance Structure Modelling

Abstract: The aim of the paper is to review three ways of modelling data quality using the covariance structure methodology, covering relations among both the latent and manifest variables – the former being the true values, while the latter are the measured data in our models. Two components of quality are presented in the models: validity and reliability, and effects of operationalisation and method effects.

The three models that have been developed in the literature are displayed: (1) The Quasi Simplex Model (QSM – Graph 1) includes a quality notion into the panel studies. (2) The Repeated Measures Model (RMM – Graph 2) covers the situation of applying several methods each repeated independently. (3) The MTMM (Multi-trait Multi-method – Graph 3) combines relations among several methods used in parallel to several concepts in one experiment – the minimal model includes three methods and three traits.

Each of the models are elaborated and briefly discussed.

Sociologický časopis, 1998, Vol. 34 (No. 2: 195-204)

Kvalita dat určuje kvalitu závěrů. Pojem kvality dat je velmi široký a také relativní. Podle typu závěrů se mohou požadavky na přesnost, spolehlivost, reprezentativitu a další aspekty velmi lišit. Budou se lišit také podle disponibilních prostředků. Některé aspekty kvality jsou nutné vždy, v některých situacích využívání informace připouštíme jisté odchylky. Nároky se mění: popis stavu v populaci (odhad procent), vyžaduje vysoký stupeň reprezentativity, konstrukce teorií, vědecká analýza a modelování kauzálních struktur vyžadují vysoce reliabilní a validní měření s nízkou chybou. Záměrem této stati je podat přehled o základních přístupech k ohodnocování měřicích instrumentů.

Základní pohled na reliabilitu a její měření je shrnut v klasické a referenční práci Lord a Novick [1968], ve které jsou definovány základní pojmy a základní vlastnosti a kde jsou také podrobně pojednány modely spolehlivosti měření pro sumační indexy. Shrnutí bylo podáno také v první části této práce Řehák [1998]. Klasická teorie je základem úvah a východiskem i pro další přístupy, má však značné limity a je různými směry rozšiřována. Rozšíření a obecnější modely se snaží především o separaci dvou základních pojmů, *reliability* a *validity*, a o možnost validitu vůbec měřit vzhledem k tomu, že v klasické teorii je měření validity značně omezeno. Campbell a Fiske [1959] navrhli a

*) Tato práce vznikla v rámci grantu Grantové agentury ČR číslo 403/95/0964 jako vstupní studie problému.

**) Veškerou korespondenci posílejte na adresu: Doc. RNDr. Jan Řehák, Sociologický ústav AV ČR, Jilská 1, 110 00 Praha 1, tel. (02) 24 22 09 79, l. 231, fax (02) 24 22 02 78, e-mail rehak@scac.cz

Andrews [1984] zrealizoval a zavedl do praxe přístup a model, který vychází z opakovaného měření téže vlastnosti několika (alespoň třemi) metodami a zahrnuje několik paralelních vlastností (alespoň tři) současně. Přístup zahrnuje oba pohledy na kvalitu dat, tj. reliabilitu i validitu a vychází ze specifické metody strukturního modelování nazývané MTMM (multitrait-multimethod) design. Ten se dnes široce využívá pro metodologické studium sociologického instrumentaria s celou řadou praktických konsekvencí. Model byl rozpracován v celé řadě dalších prací, např. ve sborníku [Saris a van Meurs 1990], ve kterém je shrnut model a problematika odhadu kvality dat pomocí MTMM před provedením první široké evropské vlny MTMM experimentů. Výsledky tohoto projektu, realizovaného v několika zemích, jsou obsahem monografie A. Scherpenzeelové [1995]. Projekt, který byl podpořen Grantovou agenturou ČR a byl u nás koordinován autorem této stati, vychází ze stejného principu a je napojením na uvedený projekt evropský. V této přehledové stati však uvedeme také, pro úplnost a možnost souběžného využití, další dva modely, které byly v literatuře rozpracovány a v praxi použity.

Měření kvality měření

V každém měření se snažíme oddělit a změřit dva základní aspekty, které jej charakterizují: *validitu* a *reliabilitu*. Zjednodušeně řečeno, validita je vlastnost zachycení záměru (měříme to, co chceme měřit) a reliabilita je vlastnost přesného měření toho, co míra opravdu obsahově odráží – přesnost bez ohledu na to, zda naše konstrukty jsou správně voleny. *Základní a nutnou podmínkou úspěšného měření, a tedy kvalitní informace, je proto validita*. Oba aspekty závisí v sociologii na volbě formy otázky, volbě slov, kategorií odpovědí, ale i na kontextu, způsobu položení otázky, resp. úplnosti proškolení osob, které zaznamenávají jevy atd. Validita sociologické otázky závisí na správné volbě z možných synonym, na sémantice celé formulace otázky a na kontextovém významu slov, na použitých kvantifikátorech při verbalizaci škál, ale i na geografické a sociální univerzalitě chápání obsahu slova.

Operacionální definice, které umožňují modelové zkoumání pojmu a empirickou charakterizaci pojmu u konkrétních měř, podávají Lord a Novick [1968]. (Viz též Řehák [1998] a Řehák et al. [1998].) Model vychází z představy existence skutečné hodnoty T , jež je konkrétně zjišťována měřením jako měřená hodnota X . Zjištěná hodnota je zatížena chybou měření:

$$(1) \quad X = T + E$$

Přitom požadujeme, aby měření proběhlo tak, že chyba E není závislá na hodnotě T .

Reliabilita měření (nebo *spolehlivost* měření) je definována jako koeficient determinace $T \rightarrow X$, $r^2(T, X)$.

Koeficient r^2 však není přímo zjistitelný, neboť sice známe nositele informace o T , kterým je X , ale skutečnou hodnotu T zjistit většinou nelze. Smyslem měření je konstruovat takový postup, aby X bylo co nejvhodnějším a co nejpřesnějším nositelem informace o neznámé a plně nezjistitelné skutečné hodnotě. I v případech, kdy je hodnota T zjistitelná, může však být měření podle rovnice (1) výhodné, neboť kompromis v akceptování jisté velikosti E může vést ke snížení nákladů, k organizační realizovatelnosti sběru informace či k jeho časovému zvýhodnění.

Teorie a pojmy, které shrnuli a dále vyvinuli Lord a Novick, umožňují empiricky odhadnout spolehlivost měření několika způsoby. Pro odhad reliability jednotlivých položek se především používá metoda „test-retest“. Pro zjištění spolehlivosti sumačních in-

dexů paralelních měření je vhodná metoda „*split-half*“. (Viz též Řehák [1998] a Řehák et al. [1998].) Klasická teorie však nabízí pouze omezené možnosti měření validity, de facto pouze korelací s vnějšími kritérii, a navíc oba pojmy se posuzují odděleně. Proto existuje celá řada pokusů model zjišťování validity zlepšit, konečného řešení však dosud nebylo dosaženo.

Situace měření subjektivních, a tím přímo neměřitelných vlastností je ale složitější, než vyjadřuje uvedená definice; lze ji vyjádřit jako:

vlastnost, kterou chceme měřit → *vlastnost, kterou měříme* → *výsledky měření*

Schéma vztahu: koncept, konstrukt, proměnná

$$F_X \quad \rightarrow \quad T_X \quad \rightarrow \quad X$$

Schéma zahrnuje dvě relace:

- Vztah T_X a X odpovídá klasické definici, je zaveden jako *reliabilita*; je závislý na metodě realizace a na některých situačních okolnostech sběru dat. Vzhledem k povaze chyby nazýváme výsledek tohoto vztahu také *efekt metody* nebo, v kontextu dotazníkových šetření, také *efektem dotazu*.
- Vztah F_X a X nazýváme *validitou* konstruktů; je závislý navíc na operacionalizaci a na obsahové vhodnosti verbalizovaných instrumentů a na dalších situačních okolnostech sběru. Závisí nejen na reliabilitě, ale řetězově také na vztahu $F_X \rightarrow T_X$, který právě tyto aspekty zahrnuje. První část řetězce tak můžeme nazvat *efektem operacionalizace* – vyjadřuje chybu, které se dopuštíme při přechodu od konceptu ke konstrukt, se kterým dále ve skutečnosti při empirické analýze dat pracujeme.

V definici reliability vystupuje jako skutečná hodnota proměnná T_X . V definici validity měření proměnná F_X .

Model lze zapsat v jeho lineárním zjednodušení jako:

$$\begin{aligned} X &= T + E \\ T &= a + bF + U \\ (2) \quad X &= a + bF + U + E \\ \text{cov}(F, U) &= \text{cov}(F, E) = \text{cov}(E, U) = 0 \end{aligned}$$

V rovnici, kde jsou všechny tři proměnné členy na pravé straně vzájemně nekorelované, je:

X = výsledek měření,

F = koncept, vlastnost, kterou zamýšlíme měřit,

T = konstrukt, vlastnost, kterou měříme operacionalizovanými měřicími prostředky,

a = konstanta posunutí stupnice,

b = koeficient převodu stupnice skutečné hodnoty na měřicí škálu (regresní koeficient F na T , tj. převod zamýšlené na skutečně realizovanou vlastnost),

U = kontextuální a obsahové chyby operacionalizace, působí *nevalidnost* konstrukt, kterou měříme velikostí $\text{var}U$ a jejím podílem na $\text{var}T$ resp. na $\text{var}X$,

E = nevychýlená složka chyby, ovlivňující spolehlivost.

Uvedeme zde dále tři schémata, která byla vyvinuta pro měření reliability a validity a která znamenají modifikace uvedené obecné definice: (A) Quasi Simplex Model, (B) Model opakovaných násobných měření (Repeated Multimethod Model), (C) MTMM design (multitrait-multimethod). Tato schémata byla vyvinuta v různých kontextech a pro řešení různých úloh, proto se hodí vždy pro určité uspořádání sběru dat. Všechna vycházejí primárně z definic (2), i když je explicitně v plné míře nevyužívají. Kritickým členem každého modelu je *nevalidnost* U , který charakterizuje odchylku obsahu a odklon konstruktu od latentní stupnice konceptu. U je chápáno jako náhodná veličina, tedy zkreslení způsobené při sběru dat (dotazu) chybným chápáním obsahu otázek, nepřesnou indikací. V modelu není zahrnuto systematické obsahové zkreslení způsobené výběrem verbalizací a dalšími omyly na straně výzkumníka. Tedy i modely, které popisujeme, mohou řešit problém validity jen částečně.

A) Quasi Simplex Model (QSM)

QSM byl formulován Heisem [1969] pro úlohu ohodnocení instrumentů v panelových uspořádáních sběru dat a v situacích, kdy nejsou k dispozici baterie položek indikujících stejnou vlastnost či postoj. Model byl modifikován později v práci Wiley a Wiley [1970] a realizován Alwinem [1989]. Model neobsahuje složku invalidity, kterou je možno odhadovat pouze při použití více metod.

Situace sběru dat: jedna otázka je v panelovém souboru pokládána opakovaně v několika krocích, dotaz je veden stejně, je používána jedna metoda. Skutečná hodnota T se může v čase ovšem měnit, proto model obsahuje rovnici změny T mezi krokem $k-1$ a k .

Schéma vztahů modelu pro krok k a mezi následnými kroky:

$$\begin{aligned} X_k &= \lambda_k T_k + E_k \\ (3) \quad T_k &= a_k + d_k T_{k-1} + U_k \end{aligned}$$

přičemž předpokládáme nekorelovanost jednotlivých členů (T, E) , (T, U) , (E, U) , E i U mezi kroky, mezi sebou.

Index k = krok v panelovém opakování měření.

Konstanta λ_k proměnlivá pro jednotlivé kroky vyjadřuje možnost změny reliability v čase.

Vztah pro T_k vyjadřuje změnu v čase: konstanty a_k a d_k v autoregresních rovnicích vyjadřují systematickou změnu hodnot mezi kroky panelového sběru, tj. časové změny proměnné T .

Pokud se v experimentu vyskytuje více metod měření paralelně pro tytéž vlastnosti, jsou koeficienty rovnic ovšem ještě indexovány podle metod.

V grafu 1 lze dosadit ohodnocení hran podle odhadnutých parametrů rovnic (3).

Aby byly parametry tohoto modelu odhadnutelné (identifikovatelné), je nutno přidat ještě další omezení. Heise [1969] navrhl řešení, které spočívá v předpokladu standardizace všech proměnných a rovnosti všech tří λ (předpoklad, že reliabilita a velikost chybové variance se nemění v čase). Saris [1990] upozorňuje také na fakt, že model není vhodný pro odhad invalidit (k tomu je zapotřebí aplikovat více metod).

V případě tří opakování model předpokládá, že reliabilita je stejná ve všech časových bodech. Odhadem reliabilit je standardizované λ . Je-li opakování více, můžeme odhadovat spolehlivost měření v každém jednotlivém opakování zvlášť, kromě prvního a posledního, které jsou však takto neidentifikovatelné [Werts et al. 1971].

Model QS je tedy aplikovatelný pouze pro panelové studie a jeho členy vyjadřující invaliditu jsou tu vedeny doplňkově, aby zachycení vývoje v panelu bylo co nejpřesnější. Vlastní odhad invalidit se tímto postupem nedoporučuje.

Graf 1: Quasi Simplex model (trojí opakování jednoduché otázky)

B) Model opakovaných násobných měření (RMM – Repeated Multimethod Model)

Model opakovaných násobných měření vychází z principu paralelní aplikace několika metod zjišťování téhož konceptu. Princip schématu pro sběr dat je založen na několika konstruktech (operacionalizacích) konceptu, který je předmětem zájmu. Každý z konstruktů aplikujeme *nezávisle* (alespoň) dvakrát po sobě.

Graf 2: Model opakování více metod (RMM)

Při uspořádání RMM je pro jednu proměnnou nutné aplikovat alespoň dvě metody pro její měření. V grafu 3 je uvedeno schéma struktury vztahů pro dvě postupná měření každé ze tří metod pro jednu proměnnou.

Model v grafu obsahuje jednu proměnnou, kterou chceme měřit – koncept F . Ta je operacionalizována třemi konstrukty T_1, T_2, T_3 (určené třemi různými metodami měření). Každý z konstruktů je ovlivněn invaliditou U_k . Všechny tři metody jsou realizovány dvakrát s výsledky X_{k1}, X_{k2} . Tyto empirické výsledky jsou ovlivněny chybami E_{k1}, E_{k2} , které způsobují sníženou reliabilitu. Hrany grafu jsou ohodnoceny pomocí koeficientů příslušných rovnic. Model definuje reliabilitu stejně jako QSM. Čtverec standardizovaného koeficientu lambda mezi T a X je odhadem reliability stejně tak, jako je odhadem reliability koeficient korelace mezi oběma měřeními X odpovídajícími jednomu T .

Model může být rozšířen pro několik konceptů F_m .

Rovnice modelu kopírují vztahy (2) s rozlišením metod T_k . Mezi každou z T_k a jejími dvěma realizacemi předpokládáme stejnou reliabilitu (jde o repliku jedné metody). Každá z metod má vlastní reliabilitu i validitu.

Rozdíl oproti QSM je v tom, že a) aplikujeme několik metod měření pro tutéž vlastnost, b) předpokládáme nezávislé realizace každé metody, c) realizace nejsou časově proměnlivé, tj. pokud opakovaná realizace je časově posunuta, předpokládáme, že mezi nimi nedochází k časovému autoregresnímu posunu.

Model umožňuje odhad obou měř kvality a také komparovat kvalitu použitých metod. Metody mohou být buď jiné verbalizace stejného obsahu, dotazy s různými škálami odpovědí, nebo i jiné formy měření (přímá otázka, sumační indexy indikační baterie, preferenční škály, výsledky jiných modelů, např. separované dimenze analýzy sdružených měření apod.).

Opakování měření může být provedeno na začátku a na konci dotazníku, opakováním dotazu po určitém časovém intervalu nebo i přístupem paralelních měření, tj. nezávislou aplikací obsahového ekvivalentu.

C) MTMM design (multitrait-multimethod)

Přístup MTMM vychází z násobného měření několika konceptů současně, přičemž každý koncept je měřen paralelně stejnými metodami.

MTMM design vychází z poněkud jiné parametrizace zahrnující přímý vliv metody M a validitu, jako přímý vztah F k X :

$$(4) \quad X_{km} = b_{km}F_k + g_{km}M_m + E_{km}$$

Indexy k odpovídají konceptům a indexy m metodám.

M reprezentuje všechny vlivy metody na X , chyby E zahrnují jednak faktory nespolehlivosti a jednak faktory invalidity, které od sebe nemohou být separovány.

Parametry rovnic (4) charakterizují dvě stránky kvality empirického měření koeficienty: koeficienty b vyjadřují validitu měření každého z konceptů a koeficienty g vyjadřují efekt metody.

Graf 3 vyjadřuje uvedené vztahy. Zároveň zavádí případnou přirozenou korelovatnost konceptů mezi sebou.

Graf 3: MTMM design (3T3M - 3 metody M a 3 vlastnosti F)

Aby byla metoda MTMM aplikovatelná, je nutné zajistit alespoň tři metody pro tři vlastnosti (graf reprezentuje minimální model).

Parametry rovnic tedy po odhadu komparují kvalitu konkrétní administrace metody i kvalitu vztahu mezi konceptem a operacionalizovanou dotazníkovou otázkou z hlediska obsahu. Je možné porovnat jak vhodnost navržených otázek jak pro teoretickou a interpretativní práci, tak přesnost měření a vhodnost formy dotazu.

MTMM nabízí přednost v tom, že v jednom modelu separuje čistou validitu konstruktů vzhledem k vlastnosti, kterou zamýšlíme měřit, a čistý efekt metody. Další předností je relativně jednoduchá praktická rozšiřitelnost na větší počet vlastností a modifikovatelnost modelu.

Přístup ale přináší také limity a komplikace. Především je to neidentifikovatelnost invalidit U a jejich separaci v rámci celkové chyby modelové rovnice E . Také technické problémy nejsou zanedbatelné. Vysoká senzitivita odhadů na malé změny parametrů vyžaduje značné výběrové rozsahy v experimentu. Model nemá zcela jasné statistické vlastnosti, jako jsou výsledky testů shody [Saris 1990, Saris a Satorra 1988, Satorra a Saris

1985]. Největší překážkou konkrétní aplikace je však obtížnost realizace příslušných schémat v praxi, neboť specifická dotazová situace vyžaduje panelový přístup a vysoké nároky na respondenty násobným opakováním dotazu na totéž. Vyžaduje alespoň trojí dotaz na totéž ale tak, aby si respondenti nepamatovali předchozí odpovědi (eliminace efektu zapamatování) a aby nefungoval efekt učení (respondent si v průběhu dotazování uvědomuje názor nebo si jej dokonce tvoří). V experimentu je možné provést maximálně dvě formy dotazu (dvě metody) během jednoho rozhovoru (na jeho začátku a na jeho konci). Třetí dotaz musí být časově separován, a proto je panel nutný.

Přes limity a obtíže je tato metoda zajímavá tím, že jí lze používat pro ohodnocení různých forem otázek v dotazníku. Kumulované informace z aplikací a tvorba databáze koeficientů rovnic pro různé typy metod i konceptů může být použita k metaanalýze a obecnějším závěrům s vysokou metodologickou relevancí pro empirickou výzkumnou práci [viz Scherpenzeel 1995].

Závěr

Metody uvedené v této stati překračují možnosti klasické teorie měření především tím, že umožňují zahrnout do modelu validitu, resp. invaliditu konstruktů a ohodnotit výzkumné instrumenty. Jejich využití je však komplikováno jednak vysokými nároky na sběr dat, a jednak metodologickými a statistickými aspekty metody strukturního modelování. Hodnocení kvality sociologických instrumentů, ať už se týká jednotlivých formátů dotazu, nebo kalibrace standardizovaných měřících baterií, je však nutnou součástí vývoje sociologie. Jen tak může být kvalita empirického výzkumu soustavně zvyšována. Systematický sběr a postupně kumulovaná empirická informace o kvalitě různých typů dotazu z různých kontextů, její shrnutí a zpracování musí nahradit subjektivní a náhodilí *ad hoc* nepodložené soudy o vhodnosti formátu dotazníkových otázek a být zahrnuty do poznatkového fondu sociologie. Uvedené metody neznamenaají konečné řešení problému a jejich vývoj bude zřejmě podmíněn kumulovanou empirickou informací a novými přístupy k modelování. Jejich aplikace však přináší již dnes řadu důležitých závěrů. Cílem této stati je na uvedené možnosti upozornit a nabídnout tak jejich aplikaci.

JAN ŘEHÁK je vědeckým pracovníkem Sociologického ústavu AV ČR, kde se zabývá metodologií a analýzou dat.

Literatura

- Alwin, D. F. 1989. „Problems in the data estimation and interpretation of reliability of survey data.“ *Quality and Quantity* 23: 409-442.
- Andrews, F. M. 1984. „Construct validity and error components of survey measures: a structural modelling approach.“ *Public Opinion Quarterly* 18: 409-442.
- Campbell, D. I., D. W. Fiske 1959. „Convergent and discriminant validation by the multitrait-multimethod matrix.“ *Psychological Bulletin* 56: 81-105.
- Heise, D. R. 1969. „Separating reliability and stability in test-retest correlation.“ *American Sociological Review* 34, 93-101.
- Lord, F. M., M. R. Novick 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley Publ. Comp.
- Řehák, J. 1998. „Kvalita dat I. Klasický model měření reliability a jeho praktický aplikační význam.“ *Sociologický časopis* 34: 51-60.

- Řehák, J., I. Bártová, J. Hamanová, P. Havlanová 1998. „Kvalita dat III. Empirické výsledky pro vybrané míry a stupnice.“ *Sociologický časopis* 34 (připraveno k publikaci).
- Saris, W. E. 1990. „Models for evaluation of measurement instruments.“ Pp. 52-80 in *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies*, ed. by W. E. Saris and A. van Meurs. Amsterdam: North-Holland.
- Saris, W. E., A. van Meurs (eds.) 1990. *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies*. Amsterdam: North-Holland.
- Saris, W. E., A. Satorra 1988. „Characteristics of structural equation models which affect the power of the likelihood ratio test.“ Pp. 121-133 in *Sociometric Research, vol. 2: Data analysis*, ed. by W. E. Saris and J. N. Galhofer. London: MacMillan.
- Satorra, A., W. E. Saris 1985. „Power of likelihood ratio test in covariance structure analysis.“ *Psychometrika* 50: 83-90.
- Scherpenzeel, A. 1995. *A Question of Quality*. Amsterdam: Royal PTT Netherland NV.
- Werts, C. E., K. G. Jøreskog, R. L. Linn 1971. „Comment on the estimation of measurement error in panel data.“ *American Sociological Review* 35: 112-117.
- Wiley, D. E., J. A. Wiley 1970. „The estimation of measurement error in panel data.“ *American Sociological Review* 35: 112-117.