

Resultatstyring – en anbefaling fra Produktivitetskommissionen uden empirisk eller teoretisk belæg

Produktivtetskommissionen anbefaler, at der indføres resultatstyring i den offentlige sektor til at fremme produktivitetsudvikling. Denne anbefaling er uden belæg i såvel relevant metodisk viden om identifikation af kausale sammenhænge som med empiri på området.



ANDERS HEIDE
Forskningschef,
TrygFonden¹

»Mål på og styr efter resultater for borgerne« er en af Produktivitetskommissionens centrale anbefalinger for at øge den offentlige produktivitet. Kommissionen understreger, at det bør være »få mål«, som styres i resultatkontrakter. Anbefalingen er koblet sammen med en anden ide, nemlig at offentlige institutioner og deres ledelser på alle niveauer bør befries for de ifølge kommissionen omfattende krav til at følge bestemte regler, processer mv.

Jeg vil i det følgende argumentere for:

- at anbefalingen er baseret på præmisser, som er i modstrid med metodisk viden
- at offentlig service i praksis produceres under mindst syv forhold, hvilket gør det yderligere svært for anbefalingen at virke
- at anbefalingen derfor som forventet ikke har støtte i den eksisterende empiri på området

- at kommissionen endelig overser andre og formentlig bedre måder langsigtet at forbedre den offentlige produktivitet

Det skal understreges, at denne artikel ikke handler om hvorvidt resultatstyring kan give en engangsforbedring af produktivitet², men om hvorvidt det er en relevant metode til at sikre en løbende langsigtet forbedring af produktiviteten. Det er denne mere langsigtede vision baseret på feedback fra resultater til forbedring af indsats, som fortalene for resultatstyring fremhæver, og som kommissionen køber³.

Idéen om resultatstyring er i strid med metodisk viden

Grundlæggende gælder, at det kun yderst sjældent er muligt, at slutte fra en observeret afvigelse på et resultatmål til en specifik årsag i form af en ændret indsats. Hvis en organisation ikke aner om afvigelsen er skabt af en ændret indsats eller af en udefra kommende faktor, så handler den i blinde. Det er bare en anden formulering af det grundlæggende problem, at det er meget svært at finde kausale sammenhænge i observationelle

NOTE 1 Denne artikel afspejler forfatterens egne analyser og konklusioner, ikke TrygFondens.

NOTE 2 Stort set alle organisationer, som på en eller anden måde bliver kigget efter i sømmene, kan finde umiddelbare besparelser. Det er imidlertid mindre interessant som middel til at sikre langsigtet produktivitetsudvikling.

NOTE 3 Denne artikel omhandler ikke problemerne i resultatstyring fra et juridisk, økonomisk, forvaltningsmæssigt eller politologisk perspektiv. Men det er tankevækkende, at juridiske sammenhænge (i f.eks. patientklagesager) bedømmes offentligt ansatte som helt dominerede princip på hvilke procedurer, de har anvendt i en given situation, ikke på hvad resultatet var. Juridisk har man længe forstået, at der er masser af faktorer, som den offentligt ansatte ikke kan gøres ansvarlig for, der har indflydelse på resultatet. Økonomer har længe set på resultatstyring i et princip-agent perspektiv og har opstillet en række kriterier for, om varianter af resultatstyring (f.eks. akkordløn) overhovedet er hensigtsmæssigt. Det er næppe at overtolke denne litteratur at slå fast, at kravene er mange og svære at indfri indenfor netop offentlig serviceproduktion. Forvaltningsstudier påviser, at resultatstyring stort set ikke omhandler produktivitet, og at styringen i i hvert fald Danmark ofte har en temmelig blød og forhandlet form. Politologiske studier har også påvist, at politikere – på mange måder ganske rationelt - ikke lægger vægt på de resultatdata, som de præsenteres for. Så også fra disse fire perspektiver virker Produktivitetskommissionen anbefaling af resultatstyring mindre gennemtænkt.

data. Man har typisk ikke styr på fejlkilder, som kan vride ens resultater i alle tænkelige retninger. Derfor er det et område, hvor metodiske fremskridt kan udløse en Nobelpris (f.eks. Heckmann eller Sims og Sargent).

Ovenstående er elementær metodisk viden. Det er baggrunden for den store interesse, der findes for at lave kontrollerede eksperimenter. Hvis man bare kunne styre indsatser efter relevante mål og løbende blive klogere, var der simpelthen ikke brug for kontrollerede eksperimenter. Man kunne faktisk *hæve* den offentlige produktivitet på sundhedsområdet eller inden for landbrugsforskningen ved at opgive de mange kostbare kontrollerede eksperimenter, der foregår der.

Resultatstyring har syv yderligere problemer i den praktiske verden

Ovenstående fremstår sikkert som en akademisk måde at argumentere på. Mon ikke praktiske folk, gerne støttet af nogle managementkonsulenter, som gerne sælger resultatstyringsprojekter, i praksis kan få noget fornuftigt ud af resultatstyring?

Svaret er nej, for i offentlig praksis er der flere og værre problemerne end spørgsmålet om fejlkilder, der rejses i den metodiske indvending.

For det første kan man ikke antage, at der findes et eller nogle få observerbare resultatmål for den enkelte institution. I praksis har offentlige institutioner ofte *adskillige* målsætninger, som man ikke bare kan forenkle til et eller nogle få mål. Noget så tilsyneladende enkelt som en børnehave skal passe og udvikle børn på mange dimensioner, være åbne for forældrenes ideer, undgå infektioner og ulykker, opspore børn i særlig risiko for ordblindhed osv.

For det andet vil mange af disse borgerrelevante målsætninger først vise sig med stor *forsinkelse*. God sproglig stimulering i en børnehave kan forebygge læseproblemer år senere. God motorisk stimulering kan forebygge inaktivitet årtier senere.

For det tredje kan man ikke antage, at man kan måle disse resultater. På masser af relevante områder findes der ikke valide måleinstrumenter, og der foregår kun en yderst behersket udvikling og validering af måleinstrumenter for langt de fleste offentlige services.

For det fjerde kan man ikke antage, at en organisation overhovedet aner, at den har gennemført eller ændret på en indsats.

Uden for sundhedsvæsenet findes der kun få meningsfulde og operationelle definitioner på konkrete indsatser. I bedste fald har man nogle grove input-tal for f.eks. antal ansatte og gennemførte undervisningstimer eller nogle flagrende betragtninger om hvilken pædagogisk filosofi, man tror man følger. En erfaring fra interventionsforskning er, at reelle ændringer i hverdagspraksis er vanskelige overhovedet at definere, dernæst at gennemføre og endnu vanskeligere at opretholde på lang sigt. Inden for uddannelsesforskningen er man eksempelvis næsten på bar bund om, hvordan man påvirker skolelæreres konkrete undervisningsadfærd. Det er ikke tilfældigt, at et lodtrækningsstudie med en simpel intervention (videobaseret supervision af lærere) med en temmelig behersket effekt på 0,22 standardafvigelser opnåede publicering i det absolutte toptidsskrift *Science* (Allen, Pianta et al. 2011).

For det femte kan man ikke antage, at man overhovedet kan observere en afvigelse som følge af indsats. Det er ikke særlig velkendt udenfor forskerkredse, at effekten af selv meget velgennemførte, velbegrundede og allerede velafprøvede indsatser ofte er *lille*. Inden for uddannelsesforskning regner man f.eks. en effekt på 0,4 standardafvigelser i f.eks. læsning eller matematik som god. Inden for socialt arbejde og arbejdsmarkedsindsatser er man i samme situation (Lipsey and Wilson 1993). Prøver en skoleleder at indføre en ny indsats med et forventet effektstørrelse på 0,4 i to klasser og dernæst sammenligner med to andre klasser, har vedkommende i allerhøjeste fald ca. 50-50 chance for at kunne se en signifikant effekt⁴. Så skoleledere handler reelt i blinde, selv i dette meget ideelle eksempel. I udvikling af produktiviteten i den offentlige sektor kan man ikke tillade sig at være nærig. Indsatser på effektstørrelser på 0,2 bør findes og implementeres. Forsøg, hvor man leder efter effektstørrelser i denne størrelsesorden, skal typisk have omkring 400 deltagere og ligger dermed langt udover, hvad stort set en hvilken som helst dansk institution har med at gøre.

Et praktisk eksempel illustrerer dette. Adab, Rouse et al. (2002) har gennemgået resultatdata for behandling af blodprop i hjertet på 37 store britiske akuthospitaler. Her er vi heldige, at der findes en rimelig entydig resultatparameter, nemlig 30 dages overlevelsen,⁵ og at der – sammenlignet med andre patientgrupper – er et stort patientflow at lave analyser på. Placerer man hospitalerne i et statistisk proces kontrol plot er det imidlertid kun to af de 37 hospitaler, som har signifikant høj dødelighed, og et, der har signifikant lav dødelighed. Med andre ord har mere end 90 pct. af sygehusene insignifikante resultater. På trods af de meget ideelle forudsætninger, kan vi stort set

NOTE 4 Eksemplet er meget groft illustrativt, fordi beregningen (risiko for type II fejl i en to-halet t-test med en effektstørrelse på 0,4 og i alt 100 forsøgspersoner) er baseret på, at der trækkes lod mellem elever og ikke mellem klasser. Da klassekammerater typisk vil korrelere noget, skal der i praksis flere elever/klasser og en ganske kompliceret statistisk analyse til. Anvendt software er <https://www.ai-therapy.com/psychology-statistics/power-calculator>

NOTE 5 30 dages overlevelsen efter en procedure er en meget udbredt målestok på sygehuse. Men lægger man tilstrækkelig tryk på denne målestok, opstår der incitament til problematisk intensivbehandling af terminale patienter og problematisk tilbageholdenhed med overhovedet at tilbyde proceduren til skrøbelige patienter.

ikke bruge disse data til at udvikle behandlingen af blodprop i hjerter på, især ikke sammenlignet med kliniske forsøg.

For det sjette bør man ikke antage, at offentlige organisationer har nye og gode idéer til forbedringer af deres indsats, som de vil kaste sig over, hvis de frigøres af besværlige regler. Realiteten er, at der er langt imellem gode idéer. Nul-resultater er meget hyppigt outcome, hvis man prøver noget nyt. Blandt 90 offentligt finansierede eksperimenter i USA indenfor uddannelse viste 88 pct. ingen effekt. I det amerikanske arbejdsministeries lodtrækningsforsøg havde ca. 75 pct. svage eller ingen effekter. Den private sektor kender også dette. Google og Microsoft havde – uden at informere alle os, der har deltaget – allerede i 2012 gennemført over 13.000 kontrollerede eksperimenter med varianter af deres software. 80-90 pct. af disse forsøg på forbedringer havde ingen målbar effekt. I den farmaceutiske industri er det almindelige, at 50-80 pct. af de indledede positive resultater fra små eksperimenter (fase II) ikke holder, når de opskales til de mere definitive fase III studier (Baron 2013). Indenfor klinisk psykologi er det en erfaring, at de effektstørrelser, man har nået i universitetsbaserede studier, ofte halveres, når en indsats rulles ud til almindelig praksis (Weisz, Jensen-Doss et al. 2006). Hvis man ønsker at udvikle produktiviteten på et område, bør udvikling af nye praktisk realiserbare idéer ikke overlades til praktikere, men være en selvstændig og velfinansieret målsætning for forskere på et område.

For det syvende er der rige muligheder for at manipulere målinger. Produktivitetskommissionen lægger op til, at der skal tages konsekvenser af dårlig produktivitet. Der er mildt sagt problematiske erfaringer fra andre lande med denne »hard accountability« tilgang, i særdeleshed indenfor skoleverdenen. Langt de fleste praktisk realiserbare resultatmål kan manipuleres, og de gange forskere har ledt efter spor efter manipulation, har de som hovedregel fundet dem (Gerrish 2014). Kreativiteten inden for manipulation har været forbløffende stor. Man kan f.eks. fortælle eleverne, hvad de skal svare, man kan holde de svageste elever væk på prøvedagen, man kan specifikt træne børnene i testtagning, man kan belønne dem på dagen⁶, man kan strategisk placere børn i specialundervisning eller lavere klasser osv. Denne effekt er så velkendt, at den hele to gange inden for samfundsvidenskaben er blevet udnævnt til en lov-mæssighed i form af Campbells lov og Goodharts lov⁷.

Disse syv mere konkrete indvendinger er velkendte i litteraturen og udgør tilsammen en formidabel forhindring for at få noget fornuftigt ud af resultatstyring.

Der er ikke empirisk dokumentation for, at resultatstyring virker i praksis

Det burde være forholdsvis nemt at påvise i praksis, om resultatstyring virker produktivitetsfremmende. Der er masser af temmelig ens folkeskoler, jobcentre, lægepraksisser, børnehaver og buslinjer rundt i verden. Det er sammenlignet med andre interventioner forholdsvis enkelt efter lodtrækning at give halvdelen af disse institutioner en eller anden form for resultatstyring og lade den anden halvdel fortsætte uændret over tre-fire år, så man netop måler, om man får en løbende forbedring og ikke bare en engangsgevinst.

Det er ikke lykkedes for undertegnede, at finde et eneste sådant kontrolleret studie.

Produktivitetskommissionen henviser til et arbejdsnotat af Greve and Ejersbo Udateret årstal. Dette notat indeholder ingen selvstændig gennemgang af, om der skulle være empirisk belæg for at satse på resultatstyring. Greve og Ejersbo skriver dog, at: »*Ud fra empiriske undersøgelser af nationale og internationale organisationer og forskning i offentlig styring kan resultater af reformindsatsen opgøres bedre end tidligere. Der har været markante effekter af indsatsen for performance management...*« (s. 3). Gennemgår man deres referencer indeholder de kun henvisning til et større empirisk studie, nemlig Pollitt and Dan (2011). Der er tale om et tværeuropæisk studie, hvor man har ledt efter »impacts« af new public management tiltag. Kvalitetskravene til at man kommer med i dette studie er mildt sagt ikke høje. Således har kun fire ud af 518 studier et eksperimentelt design, heraf er ingen baseret på lodtrækning. Med lidt god vilje har ca. hvert fjerde studie en kvantitativ tilgang⁸. Resten er casestudier, historiske beskrivelser eller som det formuleres »mainly declarative (limited evidence)«. Fulgte man en lignende metodik, ville man uden problemer kunne nå frem til at auralæsning kurerer kræft. Så det følgende bør tages med alle mulige forbehold. Men Pollitt and Dan (2011) identificerer 57 studier i deres sample, som beskæftiger sig med brugerrelevante outcomes. Her finder man, at 44 pct. viser positive effekter af NPM reformer (en langt bredere kategori end resultatmåling), 23 pct. viser negative effekter og 33 pct. viser ingen eller usikker ændring. Måler man output er tallene lidt mere positive. Men det er uklart, hvor store ændringerne er. Og det er uklart, hvad det har kostet at høste de hævdede ændringer. Selv engangseffekten på produktiviteten er således aldeles ubeskrivet. Der er med andre ord ikke belæg for Greve og Ejersbos konklusion.

Der findes et betydeligt mere seriøst metastudie, som det retfærdigvis skal siges først er udkommet efter Produktivitets-

NOTE 6 Forfatteren husker stadig den flødebolle, som min gymnasieklasse fik af en rystende nervøs pædagogikum kandidat, der samtidigt specifikt bad nogle af os i klassen om at holde igen med vores lejlighedsvist noget studentikose udspørgen.

NOTE 7 Goodharts lov lyder: »Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.« Campbells lov: »The more any quantitative social indicator (or even some qualitative indicator) is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.« Citater taget fra Wikipedia.

NOTE 8 Der gøres i øvrigt ingen indsats for at skelne mellem studier, som indeholder analysestrategier, som muliggør identifikation af kausale sammenhænge og studier, der ikke gør dette.

kommissionens betænkning, nemlig Gerrish (2014). Gerrish søger engelsksprogede studier globalt og finder i alt 49 studier, der rapporterer kvantificerede mål. Han identificerer også den metodiske kvalitet, der i de mange tilfælde er ganske høj med enten eksperimentelt design eller avancerede matchnings teknikker. Gerrishs primære resultat er, at indførelsen af resultatstyring har en effektstørrelse på 0,03 standardafvigelser på et resultatmål – altså en meget lille positiv effekt. Nogle studier, især inden for uddannelse, har negative effekter, mens andre der bl.a. benytter sig af omfattende benchmarking, opnår effektstørrelser på ca. 0,1. Nogle af de studier, Gerrish anvender, handler om brug af omfattende data-registre indenfor politistyrker eller lovbestemte incitamentsprogrammer inden for jobtræning med et større antal målepunkter. Det er altså programmer, der står i modsætning til Produktivitetskommissionens anbefalinger om resultatkontrakter med »få mål«.

Det hører også med til historien, at Gerrish kan ikke kontrollere for »gaming«; altså at resultatet er opstået som følge af mere eller mindre bevidst manipulation af data. Gerrish nævner, at forskere der har ledt generelt ikke har haft svært ved at finde eksempler på gaming. Det gælder i flere af de projekter, som Gerrish analyserer f.eks. politisystemer (se <http://en.wikipedia.org/wiki/CompStat>). Dertil kommer, at kun 3,6 pct. af de effekter, som Gerrish undersøger⁹ er målinger af produktivitet, hvor der er taget højde for *omkostningerne* ved etablering af resultatstyring. En ædruelig tolkning af Gerrishs arbejde er, at der ikke er nogen dokumentation for effekten af resultatstyring på produktivitet, uagtet at der trods alt findes en vis seriøs litteratur på området.

Hvad kan man egentlig forvente sig af en Produktivitetskommission?

Anbefalinger fra en offentlig kommission bør være præget af mainstream faglighed. Enten ved at der allerede findes empirisk dokumentation for en anbefaling, eller at der findes stærke både teoretiske og praktiske grunde til at formode, at en anbefaling vil virke. Hvis man opstiller anbefalinger, som ligger udenfor disse kriterier, bør man som absolut minimum anføre dette og dernæst anbefale afprøvninger af høj kvalitet, *før* man går videre i større skala.

Resultatstyring er en gammel travet, der som så mange andre idéer på dette felt har skiftet navn adskillige gange (her er nogle af de engelske navne): management by objectives, balanced score card, statistical process control¹⁰, performance management eller performance governance.

Der er tale om et nærmest genialt konsulentprodukt. Man sælger først en drøm om højere performance og ledelsesmæs-

sig kontrol. Dernæst skal der bruges mange konsulenttimer på at finde de forskellige »key performance indicators«, som skal være SMARTER (Specific, Measurable, Attainable, Relevant, and Timely). Måske kan man sælge et IT-system oveni. Når projektet så endelig kører, kan konsulenten altid slippe afsted med at forklare de tvivlsomme, forsinkede og uoverskuelige datasæt og fraværende resultater med, at organisationen ikke har investeret tid og kræfter nok i projektet. De mange fejlslagne projekter indgår næppe i litteraturen for området (se ovenfor), der dermed vil have en tendens til at blive for positiv i sin vurdering af resultatstyring.

Netop fordi resultatstyring har så stærke kommercielle fortalere, påhviler der Produktivitetskommissionen en særlig pligt til at undersøge, hvad der findes af belæg for en evt. anbefaling.

Skab viden, styr efter processer og saml data uden frygt for konsekvenserne

En bedre strategi for en langsigtet højere produktivitet i den offentlige serviceproduktion består af to elementer (1) skabe mere viden om effektive indsatser (2) bæredygtig integration af viden i praksis.

Det er velkendt, at der uden for sundhedsvæsenet helt generelt mangler viden om effekterne af forskellige indsatser i den offentlige sektor. Der er dog en kraftig international bølge i gang, hvor en lang række fag¹¹ udvikler eksperimentel forskning og bliver bedre til at udnytte eksisterende registerdata. Selv nulresultater, der som nævnt er hyppige i denne forskning, er uhyre relevante, fordi de angiver veje man ikke bør gå for at opnå højere produktivitet. Her er det relevant at overveje langsigtede forskningspolitiske strategier for at fremme denne type forskning. Det er ikke en nem udfordring. Opbygning af registre, der reelt kan bruges til at spotte kausale effekter, mangler på talrige områder. Ofte er der kun få forskertalenter på et område. Eksperimentel forskning er ofte kostbar (om end der findes idéer til markant billiggørelse), tager lang tid og er i det hele taget et ret risikabelt karrierevalg for en forsker.

Det er en særlig udfordring, at en stor del af de produktivitet fremmende tiltag, der er tænkelige inden for den offentlige sektor, skal implementeres i form af ændringer i menneskelig adfærd, f.eks. hvordan en lærer underviser, eller hvordan en politistyrke opklarer indbrud. Denne type forandringer er ikke lette. Men der findes en række lovende konkrete idéer i litteraturen, f.eks. praksisnær træning med simulationsøvelser, brug af supervision, modul-baserede indsatser, beslutningsstøtteværktøjer og checklister og ikke mindst integration af måling og data i det daglige arbejde mv. Der er vel at mærke tale om data, af f.eks. børns sproglige udvikling eller indbrud i

NOTE 9 Gerrish finder 2.188 effektmålinger i de 49 studier og anvender generelt de enkelte effektmålinger som analyseenhed fremfor det enkelte studie. Hans studie er tavst omkring fordelingen af disse effektmål på de 49 studier.

NOTE 10 Som findes i helt seriøse versioner til brug i industriel masseproduktion.

NOTE 11 Herunder økonomi, der som fag er helt dominerende blandt Produktivitetskommissionens medlemmer.

et lokalområde, som anvendes umiddelbart af den enkelte offentligt ansatte og som netop ikke rapporteres noget sted som et resultatmål.

Problemet med implementering er i øvrigt ikke ukendt i andre sektorer. Den offentlige sektor kan lære meget af landbruget. Forbedringer i fodring eller gødskning i landbruget skal også adopteres af de enkelte landmænd. Samtidigt er landbruget er en af de sektorer i samfundet med den bedste produktivitetsudvikling. Det er bemærkelsesværdigt, hvor omfattende et konsulent støtte-system, der findes omkring den enkelte landmand. Der er ikke tale om smarte managementkonsulenter, men om dybt specialiserede og højtuddannede konsulenttjenester, som indsamler og tilpasser eksisterende viden til den enkelte landmands behov. I praksis følger landmanden en række

fastlagte processer eller planer, men der bliver så datamæssigt fulgt op på, om man når den forventede performance. Disse data indgår samtidigt i massive registre til videre forskning og udvikling.

Det er ikke rimeligt at forlange af Produktivitetskommissionen, at den skulle have opstillet færdige løsninger på disse implementeringsproblemer, som forskningen først i de seneste årtier er begyndt at slås med inden for offentlig serviceproduktion. Men det ville have været relevant at rejse problemstillingen. En langsigtet højere produktivitet i den offentlige serviceproduktion er afhængig af fremskridt på disse områder – og måske også afhængig af, at kommende produktivitetskommissioner simpelthen er fagligt bedre klædt på.

REFERENCER

Adab, P., A.M. Rouse, et al. (2002). »Performance league tables: the NHS deserves better.« *British Medical Journal* 324(7329): 95-98.
 Allen, J.P., R.C. Pianta, et al. (2011). »An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement.« *Science* 333(6045): 1034-1037.
 Baron, J. (2013). *Practical Evaluation Strategies for Building a Body of Proven-Effective Social Programs. Suggestions for Research and Program Funders.* J. Baron. Washington DC, Coalition for Evidence-Based Policy.
 Gerrish, E. (2014) *The Impact of Performance Management on Performance in Public*

Organizations: A Meta-Analysis. SSRN <http://ssrn.com/abstract=2500420> or <http://dx.doi.org/10.2139/ssrn.2500420>
 Greve, C. and N. Ejersbo (Udateret). *Udviklingen i styringen af den offentlige sektor - Baggrundspapir til Produktivitetskommissionen, Produktivitetskommissionen* - <http://produktivitetskommissionen.dk/media/142136/Baggrundsnotat%20af%20Greve%20og%20Ejersbo.pdf>.
 Lipsey, M.W. and D.B. Wilson (1993). »The Efficacy of Psychological, Educational and Behavioral Treatment - Confirmation From Meta-Analysis.« *American Psychologist* 48(12): 1181-1209.

Pollitt, C. and S. Dan (2011). »The Impact of the New Public Management in Europe – A Meta Analysis.« *COCOPS working paper no. 3.* http://www.cocops.eu/wp-content/uploads/2012/03/WP1_Deliverable1_Meta-analysis_Final.pdf.
 Weisz, J.R., A. Jensen-Doss, et al. (2006). »Evidence-Based Youth Psychotherapies Versus Usual Clinical Care - A Meta-Analysis of Direct Comparisons.« *American Psychologist* 61: 671-689.