Félagslega æskileg svörun

Mat á próffræðilegum eiginleikum íslenskrar þýðingar Balanced Inventory of Desirable Responding og tillaga að styttingu kvarðans

Ragnhildur Lilja Ásgeirsdóttir

Lokaverkefni til MS-gráðu
Sálfræðideild
Heilbrigðisvísindasvið

HÁSKÓLI ÍSLANDS

Félagslega æskileg svörun

Mat á próffræðilegum eiginleikum íslenskrar þýðingar Balanced Inventory of Desirable Responding og tillaga að styttingu kvarðans

Ragnhildur Lilja Ásgeirsdóttir

Lokaverkefni til MS-gráðu í sálfræði

Leiðbeinandi: Fanney Þórsdóttir

Sálfræðideild

Heilbrigðisvísindasvið Háskóla Íslands

Júní 2014

Prentun: Samskipti ehf
Reykjavík, Ísland 2014

Þakkarorð

Leiðbeinanda mínum, Dr. Fanneyju Þórsdóttur, vil ég þakka fyrir ómetanlega aðstoð, fræðilega innsýn og leiðsögn við gerð þessa verkefnis. Einnig vil ég þakka Vöku Vésteinsdóttur, doktorsnema í sálfræði, fyrir hugmyndina að þessu verkefni og gott samstarf. Birki Pálmasyni og Sigurþóru Bergsdóttur vil ég þakka fyrir aðstoð við gagnasöfnun og Hlín Kristbergsdóttur fyrir aðstoð við þýðingu kvarðans. Síðast en ekki síst vil ég þakka eiginmanni mínum, Jóhannesi Þór Ágústarsyni og dætrum, Emblu og Hrafnhildi Freyju, fyrir ómetanlegan stuðning, þolinmæði og aðstoð við gerð þessa verkefnis. Fjölskyldu okkar vil ég líka þakka alla aðstoðina með dæturnar á meðan á námi mínu stóð.

Efnisyfirlit

Bls.

Kafli 4: Heimildir fyrir verkefnið í heild

Eftirfarandi meistaraverkefni samanstendur af tveimur fræðigreinum. Fyrri greinin fjallar um þýðingu og próffræðilega eiginleika mælitækis sem metur félagslega æskilega svörun, *Balanced Inventory of Desirable Responding* (BIDR; Paulhus, 1991). Seinni greinin fjallar um tillögu að styttingu BIDR. Greinarnar eru hluti af stærra rannsóknarverkefni um félagslegan æskileika (*social desirability*). Skortur hefur verið á þýddum mælitækjum á þessu sviði og var þetta verkefni hluti af því markmiði að bæta úr þeim skorti. BIDR kvarðinn er eitt þekktasta mælitækið á félagslega æskilegri svörun (Leite og Beretvas, 2005; Pauls og Crost, 2004; Stöber, Dette og Musch, 2002). Í þessum inngangi verða helstu hugtök tengd BIDR kvarðanum kynnt, ásamt því að fjallað verður um þær rannsóknir sem lýst er í fræðigreinunum tveimur.

Félagslega æskileg svörun

Félagslega æskilegri svörun (FÆS; *socially desirable responding*) má lýsa sem þeirri tilhneigingu að ýkja jákvæða eiginleika sína á sjálfsmatskvörðum og draga úr neikvæðum eiginleikum. Þessi tilhneiging hefur til að mynda valdið áhyggjum þegar túlkaðar eru niðurstöður prófa við starfsmannaráðningar. Þrátt fyrir að ýmsar rannsóknir hafi stutt notkun persónuleikaprófa í ráðningarferlinu (Barrick og Mount, 1991; Mount, Barrick og Stewart, 1998; Salgado, 2003) hafa aðrar rannsóknir bent til þess að umsækjendur hafi tilhneigingu til að svara á félagslega æskilegan hátt og niðurstöður slíkra prófa gefi því ekki rétta mynd af umsækjendum (Barrick og Mount, 1996; Ones og Viswesvaran, 1998; Zerbe og Paulhus, 1987).

Ein af þeim aðferðum sem notuð hefur verið til að leiðrétta fyrir FÆS er notkun prófa sem meta hugtakið sjálft. Í gegnum tíðina virðast próf sem meta FÆS helst hafa verið notuð á þrenna vegu til að kanna áhrif slíkrar svörunar á niðurstöður sjálfsmatskvarða í rannsóknum og fela í sér að spurningalisti sem metur félagslega æskilega hegðun er lagður fyrir á sama tíma og sjálfsmatskvarðinn (Beretvas, Meyers og Leite, 2002). Í fyrsta lagi hefur verið reiknuð fylgni milli spurningalista sem meta FÆS og annarra mælikvarða. Ef fylgnin er lág er gert ráð fyrir að FÆS hafi lítil áhrif á niðurstöður hins mælikvarðans. Í öðru lagi hafa niðurstöður prófanna tveggja verið þáttagreindar til að kanna hvort um sé að ræða tvo aðskilda þætti. Í þriðja og síðasta lagi hafa sumir

rannsakendur valið að eyða svörum þeirra sem skora hátt á spurningalistum sem meta FÆS, þar sem þeir álíta að svör þeirra við sjálfsmatskvarðanum litist of mikið af FÆS.

Þau próf sem þekktust eru við mat á FÆS eru Marlowe-Crowne kvarðinn og *Balanced Inventory of Desirable Responding* (BIDR) kvarðinn (Crowne og Marlowe, 1960; Paulhus, 1984). Helsti munur milli þessara tveggja kvarða er að Marlowe-Crowne kvarðinn metur FÆS á einni vídd en BIDR kvarðinn gerir ráð fyrir að hugtakið skiptist í tvær víddir. Samkvæmt Paulhus (1984; 1991; 2002) er annars vegar um að ræða sjálfsblekkingu (*self-deception*) og hins vegar ímyndarstjórnun (*impression management*). Bæði sjálfsblekking og ímyndarstjórnun vísa til þeirrar tilhneigingar fólks að gefa fegraða mynd af sér sem lýsir sér annað hvort með því að fólk hefur ýkta jákvæða mynd af sjálfu sér, þ.e. sjálfsblekking eða reynir viljandi að ýkja jákvæða eiginleika sína og draga úr neikvæðum eiginleikum sínum, þ.e. ímyndarstjórnun. Skilgreining á hugtakinu FÆS er þó enn fremur umdeild og ekki ljóst hvort eins, tveggja eða jafnvel þriggja þátta líkan lýsi hugtakinu best. Sjötta útgáfa BIDR-kvarðans (BIDR-6) kom út árið 1988 og inniheldur 20 atriði sem meta sjálfsblekkingu og 20 atriði sem meta ímyndarstjórnun sem svarað er á sjö punkta einvíðri stiku (1=Ekki satt, 4=Að einhverju leyti satt, 7=Mjög satt; Paulhus, 1991). Paulhus (1994) hefur bæði heimilað samfellda skorun, þar sem svör við hverri spurningu eru lögð saman og tvíkosta skorun, þar sem svargildin 6 og 7 fá gildið 1 en önnur svargildi fá gildið 0. Helmingur spurninga BIDR-6 eru neikvætt orðaðar og svör við þeim atriðum því endurkóðuð við útreikning heildarskors.

Markmið rannsóknar og helstu niðurstöður

Tilgangurinn með þessu verkefni var, eins og fyrr segir, að þýða BIDR-6 kvarðann og kanna gæði þýðingarinnar. Fjallað var um niðurstöður, bæði með tvíkosta skorun og samfelldri skorun og kannað hvort önnur aðferðin væri ákjósanlegri. Í fyrri greininni, sem rituð er á íslensku, er fjallað um BIDR-6 þýðingarferlið og niðurstöður tveggja rannsókna kynntar. Fyrri rannsóknin fjallar um próffræðilega eiginleika íslenskrar útgáfu kvarðans (N=321), þar sem meðal annars var notast við aðferðir staðfestandi þáttagreiningar og seinni rannsóknin er samantekt á niðurstöðum fengnum úr ígrunduðum viðtölum (*cognitive interviews with probing*) þar sem skilningur fólks á

atriðum kvarðans var kannaður með það að leiðarljósi að bera kennsl á vandamál tengd þýðingu kvarðans (N=20). Niðurstöður þessara rannsókna benda til þess að íslensk þýðing BIDR-6 kvarðans sé sambærileg upprunalegri útgáfu kvarðans. Engar alvarlegar athugasemdir komu fram varðandi þýðingu kvarðans og var því ekki talin ástæða til að breyta þýðingu. Við fyrri rannsóknina komu þó fram vísbendingar um að þáttabygging kvarðans væri ekki nægilega góð og nokkur atriði virtust eiga lítið sameiginlegt með öðrum atriðum. Niðurstöður tölfræðilegrar greiningar benti til þess að bæta mætti gæði kvarðans töluvert með því að sleppa ákveðnum atriðum. Þessar niðurstöður, ásamt ítarlegri skoðun á fyrri rannsóknum á BIDR-6 kvarðanum, bentu til þess að þörf væri á styttri útgáfu kvarðans. BIDR-6 kvarðinn er notaður við ýmsar aðstæður og oft samhliða öðrum mælitækjum. Rannsakendur hafa stundum brugðið á það ráð að útbúa sínar eigin styttri útgáfur af kvarðanum, sem byggjast á einfaldri tölfræðilegri greiningu (s.s. þáttahleðslum eða fylgni atriða við heildarskor) á fyrra gagnasafni í þeim tilgangi að spara tíma eða bæta þáttabyggingu kvarðans (Leite og Beretvas, 2005; Pauls og Stemmler, 2003; Stöber, Dette og Musch, 2002). Þessar styttu útgáfur hafa þó aðeins verið notaðar af rannsakendunum sjálfum og ekki hefur verið samræmi í útgáfum. Því var talin þörf á vandlega unninni tillögu að styttri útgáfu kvarðans. Með það að leiðarljósi voru framkvæmdar þrjár rannsóknir sem fjallað er um í seinni greininni, sem rituð er á ensku.

Tilgangur seinni greinarinnar var að koma með tillögu að styttingu BIDR-6 kvarðans. Fyrsta rannsóknin (N=579) fjallar um niðurstöður staðfestandi þáttagreiningar og svarferlalíkana (*item response theory*). Í annarri rannsókninni var þátttakendum tilviljanakennt skipt í tvo hópa sem fengu ólík fyrirmæli. Helmingur þátttakenda (N=258) fékk hefðbundin fyrirmæli og hinn helmingurinn (N=213) fékk þau fyrirmæli að falsa niðurstöður sínar með því að ýkja jákvæða eiginleika sína og draga úr neikvæðum eiginleikum. Fyrri rannsóknir hafa bent til þess að falsa megi niðurstöður á SB og ÍS (Holden, Starzyk, McLeod og Edwards, 2000; Paulhus, Bruce og Trapnell, 1995; Reid-Seiser og Fritzsche, 2001; Stöber, Dette og Musch, 2002) og var því dregin sú ályktun að atriði sem ekki væri hægt að falsa hefðu ekki nægilega hátt innihaldsréttmæti og ættu lítið sameiginlegt með öðrum atriðum kvarðans, til dæmis vegna þess að fólk lítur ekki á tiltekna hegðun sem félagslega óæskilega. Niðurstöður

bentu til þess að slík fyrirmæli hefðu ekki áhrif á svörun fjögurra atriða. Þriðja rannsóknin fjallar svo um niðurstöður ígrundaðra viðtala (*cognitive interviews with probing*) þar sem lögð var áhersla á innihald atriðanna. Borin voru kennsl á ýmis vandamál sem flokka má í fimm þemu: a) tvöfalda neitun, b) flóknar eða óljósar staðhæfingar, c) viðkvæmar staðhæfingar, d) staðhæfingar þar sem spurt er um tvö atriði en aðeins eitt svar leyft og e) tengsl atriða og svarkvarða. Flest atriði fengu einhverjar athugasemdir sem voru þó mismunandi alvarlegar. Niðurstöður þessara þriggja rannsókna voru svo dregnar saman og tillaga að styttri útgáfu BIDR-6 kvarðans kynnt.

Með hliðsjón af niðurstöðum þessara fimm rannsókna sem lýst er í fræðigreinunum tveimur er talið mikilvægt að rannsaka nánar próffræðilega eiginleika íslenskrar útgáfu BIDR-6. Sérstaklega er mikilvægt að kanna nánar réttmæti kvarðans. Með tilliti til niðurstaðna sem lýst er í fyrri greininni er ekki talið að mæla megi með notkun kvarðans að öllu óbreyttu, nema þá ef til vill ÍS með tvíkosta skorun. Niðurstöður fyrstu rannsóknar í seinni greininni bendir þó til þess að þáttabygging kvarðans sé betri en áður var talið, þá sérstaklega við tvíkosta skorun. Helsti áhrifavaldur þar er talinn vera stærð úrtaksins. Staðfestandi þáttagreining er viðkæm fyrir smæð úrtaks (Yu, 2002) og því líklegt að niðurstöður sem birtar eru í seinni greininni séu meira lýsandi fyrir þáttauppbyggingu kvarðans. Mátgæði mælilíkans við samfellda skorun eru einnig á mörkum þess að vera ásættanleg. Þegar litið er til áreiðanleika BIDR-6 kvarðans er hann viðunandi og mjög sambærilegur við tvíkosta og samfellda skorun. Á grundvelli niðurstaðna er því talið að þýðing sé ásættanleg og hvatt til frekari rannsókna á réttmæti íslenskrar þýðingar kvarðans. Jafnframt væri mikilvægt að framkvæma frekari rannsóknir á styttri útgáfu kvarðans. Nauðsynlegt væri að endurtaka mælingar með nýju úrtaki, ásamt því að kanna réttmæti styttu útgáfunnar samanborið við réttmæti BIDR-6.

Félagslega æskileg svörun: Íslensk þýðing og próffræðilegir eiginleikar Balanced Inventory of Desirable Responding

Ragnhildur Lilja Ásgeirsdóttir, Fanney Þórsdóttir, Vaka Vésteinsdóttir

Háskóli Íslands

Blaðsíðutitill: Félagslega æskileg svörun: BIDR-6

Ragnhildur Lilja Ásgeirsdóttir er MS í sálfræði frá sálfræðideild Háskóla Íslands. Vaka Vésteinsdóttir er Phd. nemandi við sálfræðideild Háskóla Íslands. Fanney Þórsdóttir er lektor við sálfræðideild Háskóla Íslands. Fyrirspurnum vegna greinarinnar skal beint til Ragnhildar Lilju Ásgeirsdóttur. Netfang: ragnhildurlilja@gmail.com

Útdráttur

Balanced Inventory of Desirable Responding (BIDR) er eitt mest notaða mælitækið á félagslega æskilegri svörun. Það samanstendur af tveimur undirkvörðum, Sjálfsblekkingu (SB) og Ímyndarstjórnun (ÍS), sem innihalda hvor um sig 20 fullyrðingar sem svarað er á sjö punkta kvarða (1=Ekki satt, 4=Að einhverju leyti satt, 7=Mjög satt). Tilgangur þessarar rannsóknar var að þýða BIDR kvarðann og kanna próffræðilega eiginleika íslenskrar útgáfu BIDR með staðfestandi þáttagreiningu og ígrunduðum viðtölum (*cognitive interviews with probing*). Í fyrri fasa rannsóknarinnar (N=321) er þýðingarferlinu og próffræðilegum eiginleikum lýst. Niðurstöður bentu til þess að meðaltöl og fylgni milli undirkvarða væru sambærileg þeim sem finnast í erlendum rannsóknum og áreiðanleiki mælitækisins væri viðunandi. Staðfestandi þáttagreining studdi réttmæti kvarðans þó fram hafi komið vandamál sem tengjast vissum atriðum hans. Í seinni fasa rannsóknarinnar var fjallað um niðurstöður viðtala (N=20) þar sem farið var ítarlega í íslenska þýðingu kvarðans. Niðurstöður þeirrar rannsóknar bentu til þess að ekki væru til staðar alvarleg vandamál varðandi þýðingu kvarðans.

Efnisorð: félagslega æskileg svörun, staðfestandi þáttagreining, próffræði, þýðing.

Félagslega æskileg svörun hefur löngum verið áhyggjuefni við túlkun sálfræðilegra prófa. Í starfsmannaráðningum hefur þetta til að mynda valdið hugarangri þar sem niðurstöður slíkra prófa eru hafðar til hliðsjónar við val á starfsmanni. Þrátt fyrir að ýmsar rannsóknir hafi stutt notkun persónuleikaprófa í ráðningarferlinu (Barrick og Mount, 1991; Mount, Barrick og Stewart, 1998; Salgado, 2003) hafa aðrar rannsóknir bent til þess að umsækjendur hafi tilhneigingu til að ýkja jákvæða eiginleika sína á slíkum prófum og draga úr neikvæðum eiginleikum og niðurstöður slíkra prófa gefi því ekki rétta mynd af umsækjendum (Barrick og Mount, 1996; Ones og Viswesvaran, 1998; Zerbe og Paulhus, 1987). Þessi tilhneiging hefur verið nefnd félagslega æskilega svörun (FÆS).

Reynt hefur verið að bregðast við þessari tilhneigingu fólks með ýmsum aðferðum, svo sem með því að þvinga fólk til að velja milli tveggja eða fleiri staðhæfinga sem taldar eru álíka félagslega æskilegar (*forced-choice format*), með því að spyrja einhvern nátengdan viðkomandi í stað þess að spyrja viðkomandi sjálfan eða þá með því að nota aðferðir við fyrirlögn sem bjóða upp á nafnleysi, til dæmis með því að leggja spurningalistann fyrir í gegnum tölvu (sjá t.d. Paulhus, 1991 fyrir nánari umfjöllun). Með þessum aðferðum er reynt að koma í veg fyrir eða lágmarka félagslega æskilega svörun (FÆS) en einnig hefur verið reynt að leiðrétta fyrir FÆS með notkun prófa sem meta hugtakið sjálft. Í gegnum tíðina virðast próf sem meta FÆS helst hafa verið notuð á þrenna vegu til að kanna áhrif slíkrar svörunar á niðurstöður sjálfsmatskvarða í rannsóknum sem fela í sér að spurningalisti sem metur félagslega æskilega hegðun er lagður fyrir á sama tíma og sjálfsmatskvarðinn (Beretvas, Meyers og Leite, 2002). Í fyrsta lagi hefur verið reiknuð fylgni milli spurningalista sem meta FÆS og annarra mælikvarða. Ef fylgnin er lág er gert ráð fyrir að FÆS hafi lítil áhrif á niðurstöður hins mælikvarðans. Í öðru lagi hafa niðurstöður prófanna tveggja verið þáttagreindar til að kanna hvort um sé að ræða tvo aðskilda þætti. Í þriðja og síðasta lagi hafa sumir rannsakendur valið að eyða svörum þeirra sem skora hátt á spurningalistum sem meta FÆS, þar sem þeir álíta að svör þeirra við sjálfsmatskvarðanum litist of mikið af FÆS.

Þau próf sem þekktust eru við mat á FÆS eru Marlowe-Crowne (MC) kvarðinn og *Balanced Inventory of Desirable Responding* (BIDR) kvarðinn (Crowne og Marlowe, 1960; Paulhus, 1984). Marlowe-Crowne kvarðinn er uppruninn frá árinu 1960 og hefur löngum verið algengasti kvarðinn á þessu sviði (Beretvas, Meyers og Leite, 2002; Leite og Beretvas, 2005). Tilgangurinn með þróun MC kvarðans var að hanna mælitæki sem meta átti félagslega æskilega svörun án þess að innihalda atriði sem lýstu klínískum heilkennum (*pathological implications*). MC kvarðinn gerir ráð fyrir að félagslega æskileg svörun sé einvítt hugtak (Crowne og Marlowe, 1960). Fljótlega komu fram kenningar um að hugtakinu mætti betur lýsa með tveggja þátta líkani sem Block og Wiggins nefndu *Alpha* og *Gamma* og voru síðar nánar útlistaðir af Damarin og Messick (sjá Cattell og Scheier, 1961; Edwards, Diers og Walker, 1962; Jackson og Messick, 1962; Messick, 1962; Wiggins, 1959; Wiggins, 1964 í Paulhus, 2002). Þessir þættir voru upprunalega nefndir *Alpha* og *Gamma,* þar sem *Alpha* lýsir ómeðvitaðri svarskekkju í sjálfsmatskvörðum og *Gamma* lýsir viljandi fölsun. Samkvæmt Paulhus (2002) er *Balanced Inventory of Desirable Responding* (BIDR) kvarðinn uppruninn frá tilraunum hans til að samþætta kenningar um tveggja þátta líkan FÆS og sjálfsmatskvarða Sackeim og Gur (1979) sem skiptist í tvö mælitæki, *Self-Deception Questionnaire* (SDQ) og *Other-Deception Questionnaire* (ODQ). SDQ atriðin eru spurningar sem almennt eru sannar en fólk vill síður viðurkenna, svo sem „Hefur þú einhvern tímann hugsað um að drepa einhvern?“. ODQ atriðin lýsa hegðun sem talin er félagslega æskileg en er í raun sjaldgæf, til dæmis „Ég skil aldrei eftir mig rusl“. Upprunalega voru atriði BIDR valin út frá leitandi þáttagreiningu á atriðum nokkurra kvarða sem meta FÆS, *Edwards Social Desirability* kvarðanum, MMPI-lygakvarðanum, *Social Desirability* kvarða Wiggins, MC kvarðanum, SDQ kvarðanum og ODQ kvarðanum (sjá Edwards, 1957; Meehl og Hathaway, 1946; Wiggins, 1959; Crowne og Marlowe, 1964; Sackeim og Gur, 1978 í Paulhus, 1984) sem leiddi af sér tvo meginþætti. Af þeim tíu atriðum sem hlóðu hæst á fyrri þáttinn voru fimm úr SDQ kvarðanum og fimm af þeim tíu atriðum sem hlóðu hæst á síðari þáttinn voru úr ODQ kvarðanum, sem Paulhus taldi styrkja tveggja þátta líkan hans um FÆS. Mörg þessara atriða voru síðar umorðuð, atriðum bætt við og atriði fjarlægð sem leiddi að lokum til BIDR kvarða Paulhus. Samkvæmt Paulhus (1984; 1991;

2002) er annars vegar um að ræða sjálfsblekkingu[1] (*self-deceptive enhancement*) og hins vegar ímyndarstjórnun (*impression management*). Bæði sjálfsblekking og ímyndarstjórnun vísa til þeirrar tilhneigingar fólks að gefa fegraða mynd af sér sem lýsir sér annað hvort með því að fólk hefur ýkta jákvæða mynd af sjálfu sér, það er sjálfsblekking eða með því að fólk reynir viljandi að ýkja jákvæða eiginleika sína og draga úr neikvæðum eiginleikum sínum, það er ímyndarstjórnun. Skilgreining á hugtakinu FÆS er þó enn fremur umdeild og ekki ljóst hvort eins, tveggja eða jafnvel þriggja þátta líkan lýsi hugtakinu best (Helmes og Holden, 2003).

Sjötta útgáfa BIDR-kvarðans kom út árið 1988 og inniheldur 20 atriði sem meta sjálfsblekkingu (SB) og 20 atriði sem meta ímyndarstjórnun (ÍS) sem svarað er á sjö punkta einvíðri stiku (Paulhus, 1991). Atriðunum er svarað á sjö punkta einvíðri stiku sem fá orðgildin; Ekki satt (1), Að einhverju leyti satt (4) og Mjög satt (7). Paulhus (1994) hefur bæði heimilað samfellda skorun[2], þar sem svör við hverri spurningu eru lögð saman og tvíkosta skorun, þar sem svargildin 6 og 7 fá gildið 1 en önnur svargildi fá gildið 0. Við spurningar sem eru neikvætt orðaðar fá svargildin 1 og 2 gildið 1 en önnur svargildi fá gildið 0. Helmingur spurninga BIDR-6 eru neikvætt orðaðar og svör við þeim atriðum því endurkóðuð við útreikning heildarskors. Þrátt fyrir að Paulhus (1991) hafi mælt með tvíkosta skorun út frá þeim rökum að ætlunin sé að bera kennsl á þá sem ýkja svör sín og niðurstöður staðfestandi þáttagreiningar í fyrri rannsóknum (Leite og Beretvas, 2005) stutt tvíkosta skorun, hafa aðrar rannsóknir bent til þess að samfelld skorun geti verið ákjósanlegri, með tilliti til áreiðanleika, samleitniréttmætis, fylgni við persónuleikapróf og þáttabyggingar (Cervellione, Lee og Bonanno, 2009; Stöber, Dette og Musch, 2002). Uppgefinn áreiðanleiki (alfa) prófsins frá útgefanda er frá 0,67 til 0,77 fyrir SB og frá 0,77 til 0,85 fyrir ÍS (Paulhus, 1994) en allsherjargreining Li og Bagger (2007) á 110 fyrri rannsóknum á BIDR þar sem áreiðanleikastuðull var gefinn upp gaf til kynna að meðaltalið væri 0,68 fyrir SB (staðalfrávik=0,09; spönn 0,27-0,92) og 0,74 fyrir ÍS (staðalfrávik=0,09; spönn 0,32-0,88). Ekki var fylgni milli skorunaraðferðar og

[1] Hefð hefur skapast fyrir því í erlendum ritum að nota hugtakið *self-deception* til einföldunar fyrir þennan þátt og var því ákveðið að velja íslensku þýðinguna, Sjálfsblekkingu, þrátt fyrir að það sé ekki bein þýðing á hugtakinu *self-deceptive enhancement*.

[2] Til einföldunar er fjallað um skorunina sem samfellda skorun og vísað til hefðar í erlendum rannsóknum á BIDR að fjalla um *continuous scoring*, þrátt fyrir að strangt til tekið sé skorunin ekki samfelld þar sem um sjö punkta stiku er að ræða.

áreiðanleikastuðuls. Samkvæmt viðmiðum DeVellis (2003) um að áreiðanleikastuðull kvarða skuli vera að lágmarki 0,65 fyrir notkun í rannsóknum og á hópum er áreiðanleiki BIDR-6 því viðunandi. Rannsóknir Vispoel og Tao (2013) benda til þess að endurtekningaráreiðanleiki SB sé 0,78 við tvíkosta skorun og 0,83 við samfellda skorun en fyrir ÍS sé endurtekningaráreiðanleikinn 0,83 við tvíkosta skorun og 0,86 við samfellda skorun. Niðurstöður fyrri rannsókna benda til þess að sjálfsblekking (SB) og ímyndarstjórnun (ÍS) hafi fylgni við aðrar hugsmíðar, svo sem undirkvarða NEO-FFI; taugaveiklun (SB), úthverfu (SB), samviskusemi (SB og IM) og samvinnuþýði (IM) (Costa og McCrae, 1992; Pauls og Crost, 2004; Pauls og Stemmler, 2003; Stöber o.fl., 2002).

Þar sem BIDR-6 er annað af tveimur algengustu mælitækjum á sviði félagslega æskilegrar svörunar og fyrri rannsóknir benda til þess að áreiðanleiki og réttmæti sé viðunandi er talið mikilvægt að þróa vandaða íslenska útgáfu af kvarðanum. Þrátt fyrir að sjöunda útgáfa kvarðans hafi verið gefin út árið 1998 sem nefnd hefur verið *Paulhus Deception Scales* er sú útgáfa einungis fáanleg gegn greiðslu og hefur BIDR-6 kvarðinn því verið mun meira notaður í rannsóknum (Stöber o.fl., 2002). Helsti munur á sjöttu og sjöundu útgáfu kvarðans er að svarmöguleikum var fækkað úr sjö í fimm og nokkur atriði lagfærð eða fjarlægð. Rannsóknir benda þó til að þess að sjötta og sjöunda útgáfa kvarðans séu sambærilegar (Lanyon og Carle, 2007).

Markmið rannsóknar

BIDR-6 kvarðinn hefur mikið verið notaður í rannsóknum erlendis (Li og Bagger, 2006; Li og Bagger, 2007; Stöber o.fl., 2002) en ekki er vitað til þess að unnið hafi verið að íslenskri þýðingu kvarðans. Markmið þessarar rannsóknar er því að þýða og forprófa íslenska þýðingu BIDR-6 kvarðann. Þörf er á þýddu mælitæki sem hægt er að nota á Íslandi við mat á félagslega æskilegri svörun þar sem erlendar rannsóknir hafa bent til þess að FÆS hafi áhrif á mælingar með sjálfsmatskvörðum sem notaðir eru á Íslandi (Pauls og Crost, 2004; Pauls og Stemmler, 2003; Stöber o.fl., 2002). Í fyrri fasa rannsóknarinnar var kvarðinn þýddur og próffræðilegir eiginleikar hans kannaðir við tvíkosta og samfellda skorun. Við þýðingu mælitækja er mikilvægt að vandað sé til verks og reynt að tryggja að íslensk þýðing mælitækisins sé sambærileg upprunalegri útgáfu. Þær aðferðir sem algengastar hafa verið í gegnum tíðina við þýðingar á mælitækjum

hér á landi eru annars vegar þýðing og bakþýðing, sem í sinni einföldustu mynd felur í sér að tvítyngdur aðili þýðir próf, annar tvítyngdur aðili bakþýðir yfir á upprunalega tungumálið og að lokum er sú útgáfa borin saman við upprunalegu útgáfu prófsins og hins vegar að framkæma tvær sjálfstæðar þýðingar, sem eru samræmdar og bornar undir óháðan tvítyngdan sérfræðing á sviðinu (Behling og Law, 2000; Einar Guðmundsson, 2005-2006). Báðar þessar aðferðir hafa þó verið gagnrýndar í gegnum tíðina. Bent hefur verið á að ef þýðendur eru meðvitaðir um að bakþýða eigi þýðingu þeirra, hafi þeir tilhneigingu til að velja fremur orðalag sem líklegra er að leiði til réttrar bakþýðingar, fremur en að velja það orðalag sem hentar best (Hambleton, 1993). Ef sérfræðingur sinnir bakþýðingu eru einnig líkur á því að þeir geti áttað sig á inntaki og hugtökum vegna þekkingar sinnar á efninu. Bakþýðingin gæti þannig orðið mjög sambærileg upprunalegri útgáfu þrátt fyrir að þýðingu sé ábótavant (sjá Brislin, 1970 í Behling og Law, 2000). Þegar að þýðingar eru unnar af sérfræðingum er jafnframt sú hætta að orðalag þeirra einkennist af fagmáli og er því mikilvægt að fleiri komi að þýðingarferlinu. Ein leið til að bregðast við þessu er með viðtölum (*cognitive interviews)* þar sem skilningur fólks á atriðum er kannaður og borin kennsl á vandamál tengd þýðingu (Beatty og Willis, 2007; Presser og Blair, 1994). Í síðari fasa rannsókarinnar voru tekin viðtöl þar sem skilningur viðmælenda á atriðum BIDR-6 kvarðans var kannaður, með það að leiðarljósi að bera kennsl á vandamál tengd þýðingu kvarðans. Niðurstöður þessarar rannsóknar verða notaðar til að leggja mat á gæði íslenskrar þýðingar á BIDR-6 kvarðanum.

Aðferð

Þátttakendur

Fasi eitt. Í fyrri fasa rannsóknarinnar var hentugleikaúrtak notað sem samanstóð af 321 þátttakanda. Upprunalegur fjöldi þátttakenda var 375 en 54 (14,4%) svöruðu ekki öllum atriðum og voru svör þeirra fjarlægð úr gagnasafninu. Stuðst var við hefð í erlendum rannsóknum á BIDR-6 kvarðanum við þá ákvörðun (Li og Reb, 2009) til að auðvelda samanburð við erlendar niðurstöður. Kynjahlutfallið skiptist þannig að 83,8% þátttakenda voru kvenkyns og 16,2% voru karlkyns. Þátttakendur voru fæddir á árunum 1941 til 1993 (M=1976, sf=10,45) og hafði 63,3% þátttakenda lokið við háskólagráðu.

Fasi tvö. Í seinni fasa var hentugleikaúrtak notað sem samanstóð af 20 þátttakendum. Kynjaskipting var jöfn; tíu konur og tíu karlmenn. Menntun þátttakenda var breytileg og höfðu viðmælendur ýmist lokið grunnskólaprófi, verkmenntaprófi, stúdentsprófi eða háskólaprófi. Mikil dreifing var á aldri, yngsti þátttakandinn var 21 árs en sá elsti var 58 ára. Við val á þátttakendum var reynt að stuðla að því að hafa kynjahlutfall jafnt og dreifingu sem breiðasta með tilliti til aldurs og menntunar.

Mælitæki

Uppgefin meðaltöl BIDR-6 frá útgefanda eru 7,5 (kk) og 6,8 (kvk) fyrir SB en 4,3 (kk) og 4,9 (kvk) fyrir ÍS (1994). Meðaltöl í erlendum rannsóknum þar sem notast var við upprunalega útgáfu listans fyrir SB hafa verið á bilinu 3,9 til 6,2 við tvíkosta skorun en 77,2 til 87.9 fyrir samfellda skorun og kynjamunur oft ekki verið til staðar. Erlend meðaltöl fyrir ÍS hafa verið á bilinu 4,4 til 6,5 við tvíkosta skorun en 66,3 til 79,7 við samfellda skorun (Holden o.fl., 2000; Li og Reb, 2009; Stöber o.fl., 2002; Vispoel og Tao, 2013).

Þýðing BIDR-6 kvarðans. Íslensk þýðing listans var unnin af þremur þýðendum sem hafa lokið meistaragráðu í Sálfræði eða eru við það að ljúka slíkri gráðu. Þýðendur hafa allir búið í enskumælandi löndum í tvö til tólf ár og eru tveir þeirra tvítyngdir. Við val á þýðendum var stuðst við ráðleggingar Geisinger (1994) um að ákjósanlegast sé að þýðendur hafi sérþekkingu á því sviði sem mælitækið fellur undir, séu tvítyngdir og hafi þekkingu á menningu í upprunalega heimalandi mælitækisins og þar sem nota á mælitækið. Stuðst var við ráðleggingar Alþjóðlegu prófanefndarinnar (Einar Guðmundsson, 2005-2006; International Test Commission, 2000) og Behling og Law (2000) við hönnun þýðingarferlisins. Þýðingarferlið fór þannig fram að útbúnar voru þrjár sjálfstæðar þýðingar. Þýðendur hittust svo til að samræma þýðingar, samanber ráðleggingar Geisinger (1994). Farið var yfir hverja staðhæfingu og orðalag samræmt. Í þeim tilvikum þar sem verulegur munur var til staðar á merkingu atriða voru atriðin skoðuð nánar og komist að sameiginlegri niðurstöðu um þýðingu atriðis. Sem dæmi um atriði sem erfitt reyndist að þýða má nefna atriði þrjú (*„I don´t care to know what other people really think of me“*). Ekki náðu allar tillögur að þýðingum að endurspegla merkingu fyrri hluta þessa atriðis (*„I don´t care to know“*) en á endanum náðist sátt um

þýðinguna „Ég kæri mig ekki um að vita hvaða álit aðrir hafa í raun á mér“. Atriði sex reyndist þýðendum einnig flókið („*When my emotions are aroused, it biases my thinking*“). Ekki var talið ljóst hvort um væri að ræða eingöngu neikvæðar tilfinningar eða bæði neikvæðar og jákvæðar tilfinningar. Á endanum varð þýðingin „Þegar ég er í tilfinningalegu uppnámi hugsa ég ekki skýrt“ fyrir valinu. Atriði 34 var einnig flókið í þýðingu („*I never read sexy books or magazines*“). Það sem reyndist þýðendum erfiðast varðandi þetta atriði var hvaða íslenska lýsingarorð samsvaraði enska orðinu „*sexy*“ í þessu tilviki, þá sérstaklega í styrkleika og notkun í daglegu tali. Íslenska þýðingin „Ég les aldrei bækur eða tímarit með kynþokkafullu innihaldi“ var að lokum valin. Þar sem talið var að hægt væri að skilja þetta atriði á mjög ólíkan hátt, það er með tilliti til þess hvað flokkist undir „bækur eða tímarit með kynþokkafullu innihaldi“ var ákveðið að bera þessa þýðingu sérstaklega undir 10 aðila sem fengu að sjá bæði íslenska og enska útgáfu þessa atriðis. Þeir aðilar voru beðnir um að mynda sér skoðun á því hvað þeir teldu flokkast undir „*sexy books or magazines*“ annars vegar og „bækur eða tímarit með kynþokkafullu innihaldi“ hins vegar. Flestir svöruðu þessum spurningum á sama hátt sem talin var rökstuðningur fyrir því að velja þessa þýðingu. Mikil breidd var þó í því hvað þátttakendur töldu falla undir slík rit.

Þegar að fyrstu drög að þýðingu voru tilbúin var tvítyngdur Íslendingur með BA gráðu í Sálfræði, sem starfað hefur við þýðingar og búið í enskumælandi löndum mest alla ævi, fenginn til að kanna hvort áberandi merkingarmunur væri til staðar á enskri og íslenskri útgáfu listans, sem ekki reyndist vera. Að því loknu var listinn lagður fyrir þrjá yngri viðmælendur (19-24 ára) með tilliti til orðalags. Talið var mikilvægt að fá sjónarmið yngri þátttakenda á þessu stigi. Þeir viðmælendur voru beðnir um að lesa listann yfir með tilliti til þess hversu auðskilið orðalagið væri. Einu atriði var breytt í kjölfar þess þar sem tveir viðmælendur skildu ekki orðatiltækið „að jafna metin“. Þýðingu á atriði 25 („*I sometimes try to get even rather than forgive and forget*“) var því breytt frá „Ég reyni stundum að jafna metin frekar en að fyrirgefa“ í „Ég reyni stundum að hefna mín frekar en að fyrirgefa“. Þýðendur höfðu áhyggjur af því að orðalag á íslenskri þýðingu atriðisins væri sterkara og því væri fólk tregara til að samþykkja staðhæfinguna og var það kannað sérstaklega í fasa tvö. Í lokin var íslensk og ensk

útgáfa listans borin undir aðila með doktorspróf í Sálfræði sem samþykkti þýðinguna eftir að þrjár smávægilegar lagfæringar voru gerðar á orðalagi íslensku útgáfunnar.

Framkvæmd

Fasi eitt. Í fyrri fasa var BIDR-6 kvarðinn lagður fyrir á veraldarvefnum. Þátttakendum var sendur vefpóstur með beiðni um þátttöku í rannsókninni, ásamt því að þeir voru beðnir um að senda spurningalistann áfram á aðra mögulega þátttakendur. Vegna fyrirkomulags við fyrirlögn listans var ekki mögulegt að hafa stjórn á þeim aðstæðum sem þátttakendur voru í við svörun listans né reikna út svarhlutfall en kostur þessa fyrirkomulags var að úrtakið var nokkuð fjölbreytt sem var talið vega þyngra. Lögð var áhersla á það við þátttakendur að þeim væri ekki skylt að svara öllum spurningum og hætta mætti þátttöku hvenær sem er. Engin umbun var veitt fyrir þátttöku.

Fasi tvö. Í seinni fasa voru tekin viðtöl við þátttakendur þar sem kafað var ofan í skilning þeirra á hverju atriði BIDR-6 kvarðans. Viðtölin voru hálf stöðluð en spurt var sérstaklega um atriði sem reynst höfðu flókin í þýðingu (*cognitive interviews with probing*). Aðrar aðferðir eru mögulegar við forprófun spurningalista, svo sem hefðbundin forprófun þar sem spurningalisti er lagður fyrir takmarkaðan fjölda þátttakenda, líkt og um hefðbundna fyrirlögn væri að ræða í þeim tilgangi að bera kennsl á áberandi vandamál (*conventional pretests*), atferliskóðun (*behavior coding*) eða þá að rýnihópur sérfræðinga er fenginn til að yfirfara spurningalistann (*expert panels*) (Presser og Blair, 1994). Hver aðferð hefur sína kosti og galla, svo sem varðandi kostnað, tíma eða nákvæmni niðurstaðna en talið var að ígrunduð viðtöl með ítarspurningum (*cognitive interviews with probing)* hentuðu best í þessari rannsókn þar sem markmiðið var að bera kennsl á vandamál við þýðingu atriðanna og fá sem ítarlegastar upplýsingar um skilning ólíkra þátttakenda á atriðunum. Flokka má ítarspurningar í fjóra flokka (Beatty og Willis, 2007), það er fyrirsjáanlegar (*anticipated;* fyrirfram ákveðnar af rannsakanda), ófyrirsjáanlegar (*spontaneous;* ákveðnar á staðnum af rannsakanda), tilfallandi (*emergent;* ákveðnar á staðnum undir áhrifum frá svaranda) og skilyrtar (*conditional;* fyrirframákveðnar en háðar svaranda). Fjórir rannsakendur komu að undirbúningi viðtalanna og samningu handrits sem stuðst var við í viðtölunum. Ráðleggingum Fowler (1995) var fylgt við framkvæmd viðtala. Byrjað var á opnum

spurningum um atriði og hvernig viðkomandi hefði komist að niðurstöðu varðandi val á svarmöguleika. Því næst var farið ítarlega í ákveðin atriði ef viðkomandi hafði ekki þegar minnst á þau, svo sem varðandi skilning viðkomandi á ákveðnum orðum eða orðasamböndum. Í lokin voru þátttakendur enn fremur spurðir að því hvort þeir teldu atriðið hafa verið óljóst að einhverju leyti eða hvort orða mætti atriðið betur.

Viðtölin fóru fram á heimili rannsakanda, heimili viðmælanda eða vinnustað viðmælanda. Til að koma í veg fyrir truflun fóru viðtölin fram í lokuðu eða aðskildu herbergi þegar þáttakendur höfðu nægan tíma. Þar sem listinn samanstendur af 40 staðhæfingum var ákveðið að viðtölin færu fram í tveimum lotum. Í fyrra viðtalinu var farið yfir atriði sjálfsblekkingar og í seinna viðtalinu var farið yfir atriði ímyndarstjórnunar. Að meðaltali tók hvert viðtal um tvær klukkustundir í framkvæmd og því samtals um fjórar klukkustundir fyrir hvern þátttakanda. Í upphafi hvers viðtals voru viðmælendur beðnir um að lesa hefðbundin fyrirmæli sem fylgja BIDR-6 kvarðanum og útskýrt fyrir þeim að tilgangur þessarar rannsóknar væri einungis að varpa ljósi á skilning viðmælenda á atriðum listans. Því næst fengu þátttakendur tíma og næði til að svara listanum. Að lokum fór rannsakandi ásamt viðmælanda yfir hvert og eitt atriði kvarðans. Viðtölin voru hljóðrituð og síðar skrifuð upp orðrétt. Unnið var úr þeim gögnum til að bera kennsl á vandamál sem gætu verið til staðar við þýðingu atriðanna. Þátttakendur voru beðnir um leyfi fyrir upptökunum gegn því loforði að gögnum yrði eytt að rannsókn lokinni.

Tölfræðileg úrvinnsla

Við tölfræðilega úrvinnslu í fasa eitt var notast við tölfræðiforritin SPSS 20 og LISREL 9.10 (Jöreskog og Sörbom, 2012). SPSS 20 var notað við útreikning á lýsandi tölfræði og atriðagreiningu en staðfestandi þáttagreining var gerð með LISREL í þeim tilgangi að kanna þáttabyggingu íslenskrar útgáfu BIDR. Þar sem breytur voru á raðkvarða og skekktar (*Shapiro-Wilk* próf voru marktæk) var aðferð minnstu veginna ferninga, þar sem leiðrétt er fyrir smæð úrtaks (*diagonally weighted least squares*), notuð til að greina raðkvarðafylgnifylki (*polychoric correlation; tetrachoric fylgnifylki fyrir tvíkosta skorun*) og aðfellusamdreifnifylki (*asymptotic covariance matrix*) (Flora og Curran, 2004). Með staðfestandi þáttagreiningu er kannað hve vel gögnin falla að þeim

mælilíkönum sem sett eru fram. Þannig er hægt að kanna gæði beggja þátta BIDR-6 kvarðans. Til að skoða heildarmátgæði líkana er mikilvægt að styðjast við fleiri en einn mátgæðastuðul (Raykov og Marcoulides, 2006; Jöreskog, 1993). Í þessari rannsókn verður stuðst við niðurstöður Satorra-Bentler kí-kvaðratprófs, RMSEA, CFI og NNFI samkvæmt ábendingum frá Hu og Bentler (1998) og Yu (2002).

Satorra-Bentler kí-kvaðratprófið er notað í stað hefðbundins kí-kvaðratprófs þegar að aðferð minnstu veginna ferninga (*diagonally weighted least squares*) er notuð og leiðréttir sá stuðull fyrir skekkju í dreifingu (Yu, 2002). Ef niðurstöður kí-kvaðratprófs eru háar og stuðullinn marktækur (p>0,05) er núlltilgátunni hafnað sem bendir til þess að mátgæði líkansins séu ófullnægjandi en ef niðurstöðurnar eru lágar og stuðulinn ómarktækur (p<0,05) eru rök fyrir því að ekki sé hægt að hafna núlltilgátunni og mátgæðin séu góð (Albright og Park, 2009; West, Taylor og Wu, 2012). Helsti ókostur við kí-kvaðrat stuðulinn er sá að stuðullinn er háður úrtaksstærð og er því líklegur til þess að vera marktækur ef úrtak er miðlungsstórt, óháð mátgæðunum (Hu og Bentler, 1999; Raykov og Marcoulides, 2006; West, Taylor og Wu, 2012). Í minni úrtökum, N≤250, getur verið varhugavert að líta eingöngu á niðurstöður SB kí-kvaðratprófsins og er því einnig mælt með notkun RMSEA, CFI og NNFI í þeim tilvikum (Yu, 2002). Í þessari rannsókn er úrtakið aðeins lítillega yfir því viðmiði og verður því einnig litið til niðurstaðna RMSEA, CFI og NNFI við mat á gæðum líkana. RMSEA stuðullinn hefur þann kost að vera nokkuð ónæmur fyrir úrtaksstærð og segir til um það hversu illa líkanið fellur að gögnunum (sjá Bentler, 1990 í Raykov og Marcoulides, 2006). CFI stuðullinn lýsir því hversu mikið betri mátgæði mælilíkansins eru samanborið við mælilíkan sem gerir ekki ráð fyrir sambandi milli breytanna. NNFI stuðullinn ber saman mátgæði mælilíkans við mátgæði líkans sem gerir ekki ráð fyrir sambandi milli mælibreyta og leiðréttir fyrir stærð líkans. Gildi NNFI falla oftast á bilinu núll til einn en stuðullinn getur gefið af sér gildi sem falla fyrir utan þá spönn. Því hærri sem stuðullinn er því betri eru mátgæði mælilíkansins. Deildar meiningar hafa verið um viðmiðunargildi þessara mátstuðla (Markland, 2007) en almennt er fallist á að viðmiðunargildi > .90 fyrir NNFI og CFI og lægra en .08 fyrir RMSEA bendi til þess að líkan falli vel að gögnunum (Bentler, 1990; Browne og Cudeck, 1993; Hu og Bentler, 1998, 1999; Steiger, 1990; Yu, 2002; Yuan, 2005).

Niðurstöður

Fasi eitt

Lýsandi tölfræði. Líkt og sjá má í 1. töflu voru meðaltöl sjálfsblekkingar (SB) í íslenskri útgáfu BIDR-6 kvarðans sambærileg meðaltölum úr erlendum rannsóknum en meðaltöl ímyndarstjórnunar (ÍS) voru örlítið hærri en í erlendum rannsóknum (Holden, Starzyk, McLeod og Edwards, 2000; Li og Reb, 2009; Stöber, Dette og Musch, 2002).

1. tafla. Lýsandi tölfræði fyrir undirþætti BIDR-6 kvarðans við tvíkosta og samfellda skorun (n=321).

		Meðaltal	*Staðalfrávik*	*Skekkja*	*Ris dreifingar*
Sjálfsblekking	Tvíkosta skorun	5,08	3,23	0,70	0,04
	Samfelld skorun	83,59	12,08	0,05	-0,22
Ímyndarstjórnun	Tvíkosta skorun	6,83	3,76	0,43	-0,37
	Samfelld skorun	81,56	16,03	0,03	-0,02

Ekki var til staðar kynjamunur á meðaltölum SB og ÍS, hvorki með tvíkosta skorun ($t(313)=.706$, $p=.481$) fyrir SB og ($t(313)=-.133$, $p=.894$) fyrir ÍS né samfelldri skorun ($t(313)=.674$, $p=.501$) fyrir SB og ($t(313)=-.215$, $p=.830$) fyrir ÍS. Dreifingin er fremur jákvætt skekkt við tvíkosta skorun bæði fyrir SB og ÍS en víkur aðeins lítillega frá normaldreifingu við samfellda skorun. Neikvætt ris dreifingar undirþátta við tvíkosta og samfellda skorun merkir að dreifingin er örlítið lægri og halar dreifingarinnar lengri en í normaldreifingu.

Fylgni milli undirþátta BIDR kvarðans, SB og ÍS, má sjá í 2. töflu. Fylgnin er sambærileg fylgni milli undirþátta í erlendum rannsóknum (Stöber o.fl., 2002) og er hæst milli ÍS við tvíkosta og samfellda skorun og lægst milli SB við tvíkosta skorun og ÍS við samfellda skorun. Fylgnistuðlar fyrir sömu hugsmíð með ólíkum skorunaraðferðum eru í báðum tilfellum háir, 0,88 fyrir ÍS við tvíkosta og samfellda skorun og 0,78 fyrir SB við tvíkosta og samfellda skorun.

2. tafla. Fylgni milli undirþátta BIDR-6 við tvíkosta og samfellda skorun.

		Sjálfsblekking tvíkosta	*Sjálfsblekking samfelld*	*Ímyndarstjórnun tvíkosta*
Sjálfsblekking	Tvíkosta skorun			
	Samfelld skorun	0,78*		
Ímyndarstjórnun	Tvíkosta skorun	0,31*	0,32*	
	Samfelld skorun	0,22*	0,36*	0,88*

*p<,001

Próffræðilegir eiginleikar BIDR-6. Í 3. töflu má sjá áreiðanleika (alfa) undirþátta BIDR-6 kvarðans. Líkt og sjá má er hann viðunandi (DeVellis, 2003) fyrir Sjálfsblekkingu (SB) og Ímyndarstjórnun (ÍS) bæði við tvíkosta og samfellda skorun. Áreiðanleikastuðlar eru hærri fyrir ÍS en SB og örlítið hærri við samfellda skorun beggja kvarða. Áreiðanleikastuðlar eru sambærilegir áreiðanleikastuðlum úr erlendum rannsóknum og nálægt meðaltali allsherjargreiningar Li og Bagger (2007), þar sem meðaltalið var 0,68 fyrir SB og 0,74 fyrir ÍS.

3. tafla. Áreiðanleiki undirþátta BIDR-6 kvarðans við tvíkosta og samfelldri skorun.

		n	*Alfa*
Sjálfsblekking	Tvíkosta skorun	321	0,70
	Samfelld skorun	321	0,71
Ímyndarstjórnun	Tvíkosta skorun	321	0,76
	Samfelld skorun	321	0,77

Réttmæti íslenskrar þýðingar BIDR-6 kvarðans var metið með staðfestandi þáttagreiningu. Borin voru saman tvö líkön þar sem dreifni þátta var fest í einum, annars vegar líkan fyrir SB og hins vegar líkan fyrir ÍS. Algengt er að BIDR-6 kvarðinn sé ekki lagður fyrir í heild sinni og aðeins annar hvor kvarðinn notaður og eru það helstu rökin fyrir að tvö líkön voru metin. Niðurstöður staðfestandi þáttagreiningar fyrir SB og ÍS, annars vegar við tvíkosta skorun og hins vegar við samfellda skorun, má sjá í 4. töflu.

4. tafla. Niðurstöður staðfestandi þáttagreiningar við tvíkosta skorun og samfellda skorun.

		SBχ^2	*df*	*RMSEA*	*Öryggisbil RMSEA (90%)*	*CFI*	*NNFI*
SB	Tvíkosta skorun	517,66*	170	0,13	0,13-0,14	0,90	0,88
	Samfelld skorun	397,14*	170	0,15	0,14-0,15	0,84	0,83
ÍS	Tvíkosta skorun	393,20*	170	0,09	0,09-0,10	0,95	0,94
	Samfelld skorun	397,74*	170	0,12	0,11-0,12	0,90	0,88

*p<0,01

Þegar atriði SB kvarðans voru greind með tvíkosta skorun benti CFI stuðullinn til þess að mátgæði eins þátta líkans væru ásættanleg og NNFI stuðulinn var nálægt viðmiðunargildi sínu. Samfelld skorun atriða kom verr út því allir mátstuðlar bentu til þess að eins þátta líkan félli illa að gögnunum í því tilfelli. Stöðluð leif benti til þess að of há fylgni væri á milli mælivillna atriða 3, 16 og 19 en þessi atriði fjalla öll um viðhorf svarenda til álits annarra á þeim sjálfum. Einnig var há fylgni á milli mælivillna atriða 15 og 17 en þau fjalla bæði um skynsemi og trú á eigin dómgreind.

Í 5. töflu má sjá atriðin 40 í íslenskri þýðingu og staðlaðar þáttahleðslur hvers atriðis við tvíkosta og samfellda skorun. Meðaltöl og staðalfrávik atriða eru einnig birt. Þáttahleðslur fyrir SB við tvíkosta skorun eru á bilinu 0,18 til 0,67. Allar þáttahleðslur voru tölfræðilega marktækar en hleðslur þriggja atriða (3, 7 og 13) voru lægri en 0,3, sem túlka má sem lágar (Sharma, Mukherjee, Kumar og Dillon, 2005; Shevlin og Miles, 1998). Við samfellda skorun eru þáttahleðslur frá 0,00 til 0,61. Tvær þáttahleðslur voru ekki marktækar, atriði þrjú og sjö og hleðslur fimm atriða voru lægri en 0,3. Þetta voru atriði 1, 4, 8, 13 og 14. Í öllum tilfellum nema þremur eru þáttahleðslur hærri fyrir tvíkosta skorun en samfellda skorun.

Þegar atriði ÍS kvarðans voru þáttagreind með samfelldri skorun benti CFI til þess að eins þátta líkan félli vel að gögnunum og NNFI stuðullinn var nálægt viðmiðunargildi sínu. Tvíkosta skorun kom einnig betur út hjá ÍS kvarðanum og bentu CFI og NNFI mátstuðlarnir til þess að mátgæði eins þátta líkans væru ásættanleg. RMSEA stuðullinn var einnig nálægt viðmiðunargildi sínu. Stöðluð leif benti til að há fylgni væri á milli mælivillna atriða 24, 27 og 40 sem öll fjalla um tal sem oftast er talið neikvætt, það er að blóta, baktala og slúðra. Há fylgni var einnig á milli mælivillna atriða 26 og 28 sem

bæði fjalla um að taka eða skemma hluti í eigu annarra. Þáttahleðslur fyrir ÍS voru frá 0,31 til 0,71 við tvíkosta skorun. Við samfellda skorun voru þáttahleðslur frá 0,26 til 0,61. Allar þáttahleðslur voru marktækar en eitt atriði var undir 0,3 við samfellda skorun, atriði 34.

5. tafla. Þáttahleðslur, meðaltöl og staðalfrávik atriða í staðfestandi þáttagreiningu við tvíkosta skorun (T) og samfelldri skorun (S).

	Þáttahleðslur		*Meðaltal*	
	T	*S*	*T*	*S*
Sjálfsblekking				
1. Það álit sem ég mynda mér á fólki við fyrstu kynni reynist yfirleitt rétt.	0,30	0,26	0,19(0,39)	4,44(1,18)
2. Það væri erfitt fyrir mig að losa mig við einhvern af mínum ósiðum.*	0,45	0,43	0,23(0,42)	4,27(1,46)
3. Ég kæri mig ekki um að vita hvaða álit aðrir hafa í raun á mér.	0,18	0,03[a]	0,09(0,29)	3,27(1,45)
4. Ég hef ekki alltaf verið heiðarleg/ur gagnvart sjálfri/sjálfum mér. *	0,34	0,21	0,17(0,37)	3,73(1,67)
5. Ég veit alltaf af hverju mér líkar við eitthvað.	0,66	0,46	0,27(0,45)	4,45(1,50)
6. Þegar ég er í tilfinningalegu uppnámi hugsa ég ekki skýrt. *	0,67	0,57	0,12(0,33)	3,53(1,57)
7. Þegar ég hef gert upp hug minn geta aðrir sjaldan fengið mig til að skipta um skoðun.	0,23	-0,01[a]	0,12(0,32)	3,56(1,51)
8. Ég er ekki öruggur bílstjóri þegar ég keyri yfir hámarkshraða. *	0,35	0,20	0,21(0,41)	3,76(1,81)
9. Ég hef fullkomna stjórn á eigin örlögum.	0,32	0,39	0,15(0,36)	3,69(1,59)
10. Það er erfitt fyrir mig að bægja frá mér óþægilegum hugsunum. *	0,65	0,60	0,26(0,44)	4,11(1,69)
11. Ég sé aldrei eftir ákvörðunum mínum.	0,45	0,38	0,06(0,24)	2,90(1,40)
12. Ég missi stundum af tækifærum því að ég er ekki nógu fljót/ur að gera upp hug minn. *	0,44	0,37	0,22(0,42)	4,06(1,62)
13. Ég kýs vegna þess að atkvæði mitt getur skipt sköpum.	0,26	0,16	0,58(0,49)	5,47(1,52)
14. Foreldrar mínir voru ekki alltaf sanngjarnir þegar þeir refsuðu mér. *	0,34	0,28	0,39(0,49)	4,69(1,81)
15. Ég er fullkomlega skynsöm manneskja.	0,61	0,45	0,30(0,46)	4,53(1,52)
16. Ég kann sjaldnast að meta gagnrýni. *	0,51	0,34	0,37(0,48)	4,92(1,32)
17. Ég hef mikla trú á dómgreind minni.	0,66	0,46	0,49(0,50)	5,31(1,13)
18. Ég hef stundum efast um hæfni mína sem elskhugi. *	0,43	0,46	0,29(0,46)	4,17(1,77)
19. Það er allt í lagi mín vegna þó einhverju fólk kunni að líka illa við mig.	0,40	0,31	0,34(0,48)	4,66(1,63)
20. Ég veit ekki alltaf af hverju ég geri það sem ég geri. *	0,59	0,61	0,23(0,42)	4,07(1,57)
Ímyndarstjórnun				
21. Ég lýg stundum ef þörf krefur. *	0,71	0,61	0,31(0,46)	4,36(1,63)

22. Ég hylmi aldrei yfir mistök mín.	0,64	0,43	0,26(0,44)	4,09(1,66)
23. Það hefur komið fyrir að ég hef notfært mér einhvern. *	0,65	0,60	0,33(0,47)	4,58(1,55)
24. Ég blóta aldrei.	0,37	0,35	0,08(0,28)	2,23(1,67)
25. Ég reyni stundum að hefna mín frekar en að fyrirgefa. *	0,55	0,54	0,63(0,48)	5,56(1,45)
26. Ég fylgi alltaf lögum þó ólíklegt sé að það komist upp um mig.	0,41	0,41	0,50(0,50)	4,95(1,67)
27. Ég hef baktalað vin/vinkonu. *	0,69	0,51	0,22(0,42)	3,75(1,75)
28. Þegar ég heyri fólk tala saman í trúnaði forðast ég að hlusta á samræðurnar.	0,53	0,37	0,36(0,48)	4,60(1,70)
29. Ég hef fengið of mikinn pening til baka í verslun án þess að segja afgreiðslufólkinu frá því. *	0,36	0,32	0,46(0,50)	4,65(2,18)
30. Ég tel alltaf fram allan tollskyldan varning.	0,54	0,42	0,20(0,40)	3,15(2,09)
31. Þegar ég var ung/ur stal ég stundum. *	0,38	0,40	0,50(0,50)	4,79(2,19)
32. Ég hef aldrei skilið eftir mig rusl á götum.	0,38	0,34	0,25(0,43)	3,27(2,15)
33. Ég ek stundum yfir hámarkshraða. *	0,46	0,43	0,11(0,31)	2,74(1,72)
34. Ég les aldrei bækur eða tímarit með kynþokkafullu innihaldi.	0,31	0,26	0,14(0,35)	2,83(1,88)
35. Ég hef gert hluti sem ég segi ekki öðru fólki frá. *	0,32	0,31	0,15(0,35)	2,73(1,89)
36. Ég tek aldrei hluti sem ég á ekki.	0,48	0,33	0,59(0,49)	5,16(1,96)
37. Ég hef tekið mér veikindaleyfi frá vinnu eða skóla þrátt fyrir að vera ekki veik/ur. *	0,46	0,42	0,46(0,50)	4,39(2,31)
38. Ég hef aldrei skemmt bókasafnsbók eða vöru í verslun án þess að láta vita af því.	0,62	0,43	0,72(0,45)	5,60(1,93)
39. Sumar venjur mínar eru mjög slæmar. *	0,56	0,48	0,26(0,44)	3,93(1,76)
40. Ég slúðra ekki um mál annarra.	0,59	0,43	0,30(0,46)	4,19(1,70)

* Svör við þessum atriðum eru endurkóðuð við tölfræðilega úrvinnslu, ª Þáttahleðsla ekki marktæk (>0,05), T = tvíkosta skorun, S = samfelld skorun, gildi í sviga eru staðalfrávik.

Fasi tvö

Niðurstöður viðtala bentu til þess að mögulega mætti bæta þýðingu atriða 13, 16, 19, 23, 25 og 39 og hugsanlega er orðalag atriða 13, 23, 25 og 39 of sterkt. Hafa ber þó í huga að mörg atriðanna eru viljandi mjög sterkt orðuð í upprunalegri útgáfu kvarðans þar sem ætlunin er að bera kennsl á ýkta svörun en mikilvægt er að bera kennsl á atriði sem mögulega geta verið orðuð á enn sterkari hátt í íslenskri þýðingu.

Orðið „*meta*“ í atriði 16 „*Ég kann sjaldnast að meta gagnrýni*“ vafðist fyrir einhverjum þátttakendum. Viðmælendur lögðu tvenns konar skilning á orðið, annars vegar skildu þeir hugtakið á þann veg að átt væri við hvort þeir leggðu mat á gagnrýni og tækju gagnrýni til umhugsunar og hins vegar hvort þeir væru ánægðir með að fá gagnrýni. Mögulegt er að þetta megi rekja til vandamáls við þýðingu en jafnframt gæti verið að sama vandamál sé til staðar í upprunalegri útgáfu kvarðans. Orðið „*appreciate*“

er þar notað en túlka má það orð á samskonar hátt. Flestir töldu fullyrðinguna einungis eiga við um gagnrýni á þá sjálfa en einn viðmælandi taldi að einnig væri átt við gagnrýni í garð annarra. Þrátt fyrir að þátttakendur hafi ekki talið sig lenda í vandræðum með atriði 19 „*Það er í lagi mín vegna þó einhverju fólki kunni að líka illa við mig*“ lögðu ekki allir sama skilning á hugtakið „*einhverju fólki (some people)*“. Hluti viðmælenda taldi að eingöngu væri átt við ókunnuga en aðrir töldu að þeir sem væru nákomnir gætu einnig fallið undir þennan hóp. Nokkrir viðmælendur töldu orðið „*sumir*“ eiga betur við í þessu tilviki þar sem þeir töldu að með því hugtaki væri bæði átt við ókunnuga og þá sem stæðu þeim nær. Aðrir viðmælendur töldu ekki skipta máli hvort orðið væri notað og töldu orðin hafa sömu merkingu. Talið er að þetta vandamál megi þó frekar rekja til upprunalegrar útgáfu atriðisins fremur en þýðingu. Orðið „*notfært (taken advantage)*“ í atriði 23 „*Það hefur komið fyrir að ég hef notfært mér aðra*“ var túlkað á ólíkan hátt meðal þátttakenda. Sumir viðmælendur töldu orðið vísa til þess að nýta sér aðstoð annarra en aðrið töldu að orðið hefði neikvæðari merkingu og hefði slæmar afleiðingar fyrir hinn aðilann. Þeir þátttakendur sem töldu að orðið hefði neikvæða merkingu töldu atriðið of sterkt orðað. Atriði 25 „*Ég reyni stundum að hefna mín frekar en að fyrirgefa*“ var túlkað á ólíkan hátt meðal þátttakenda. Í fyrsta lagi töldu þátttakendur orðið „*hefna*“ eiga við um misalvarlega atburði. Sumir töldu orðið eiga við um mjög alvarlega hegðun á meðan að aðrir tengdu orðið við léttvægari hegðun, svo sem að hefna sín í fótbolta. Í öðru lagi var breytilegt hvernig þátttakendur túlkuðu orðið „*fyrirgefa*“ en sumir töldu sig vera búna að fyrirgefa þegar þeir hættu að hugsa um atburðinn en aðrir töldu fyrirgefningu fela í sér að ræða við viðkomandi aðila. Þátttakendur voru því ekki alveg sammála um skilgreiningu hefndar og fyrirgefningar. Við þýðingu þessa atriðis var ákveðið að þýða atriðið „*I sometimes try to get even rather than forgive and forget*“ ekki beint þar sem það þótti ekki nægilega auðskiljanlegt á íslensku máli. Þýðendur voru meðvitaðir um að íslensk þýðing atriðisins væri ef til vill of sterkt orðuð og er því mikilvægt að líta til niðurstöðu tölfræðilegar greiningar við ákvörðun á því hvort breyta þurfi þýðingu þessa atriðis. Sumir þátttakendur töldu atriði 39 „*Sumar venjur mínar eru mjög slæmar*“ mögulega vera of sterkt orðað og töldu að fjarlægja þyrfti orðið „*mjög*“ þar sem það gæfi til kynna venjur sem hefðu mikil áhrif á líf og heilsu viðkomandi. Aðrir töldu sig ekki lenda í vandræðum með þessa fullyrðingu og nefndu mörg dæmi um

slíkar venjur, svo sem reykingar, hreyfingarleysi, áfengisvandamál, óstundvísi, kæki og fleira. Hafa ber þó í huga að í upprunalegri útgáfu kvarðans er fjallað um „*pretty awful habits*“ þannig að ekki er talið að íslensk þýðing sé of sterkt orðuð. Atriði 13 „*Ég kýs vegna þess að atkvæði mitt getur skipt sköpum*“ var túlkað af flestum viðmælendum á þann veg að atkvæði skipti sköpum ef það ræður úrslitum kosninga en nokkrir töldu atkvæðið ekki þurfa að ráða úrslitum. Sumir viðmælendur töldu að atriðið væri það sterkt orðað að erfitt væri að svara staðhæfingunni á þann veg að hún ætti að öllu leyti við um viðkomandi. Upprunalega útgáfan notar enska orðalagið „*make a difference*“ og er því mögulegt að íslenska þýðingin sé of sterkt orðuð. Töldu þessir viðmælendur að orða þyrfti þessa staðhæfingu á vægari hátt til að fá meiri dreifingu á svörum. Atriði 13 var eitt þeirra atriða sem var erfitt í þýðingu.

Álitaefni höfðu einnig komið upp við þýðingu annarra atriða og voru þau skoðuð sérstaklega í viðtölunum. Þetta voru atriði 3, 6 og 24. Niðurstöður viðtala veittu frekari upplýsingar um skilning þátttakenda á þeim atriðum. Atriði 3, „*Ég kæri mig ekki um að vita hvaða álit aðrir hafa í raun á mér*“, reyndist nokkuð flókið í þýðingu. Nokkrir þátttakendur töldu orðið „*vita*“ hafa áhrif á skilning þeirra á atriðinu, þar sem þeir vildu síður vita álit annarra þrátt fyrir að hafa áhyggjur af áliti þeirra. Enska orðið „*know*“ er þó notað í upprunalegri útgáfu listans þannig að ekki er talið að það vandamál tengist þýðingu kvarðans. Alvarlegasta vandamálið við þetta atriði virðist þó stafa af tvöfaldri neitun sem einnig er til staðar í upprunalegri útgáfu kvarðans. Af tuttugu þátttakendum, svöruðu sjö þátttakendur atriðinu öfugt. Talið er að þetta vandamál sé til staðar í fleiri atriðum í kvarðanum og er því þörf á að skoða nánar vandamál sem kunna að stafa af innihaldi atriðanna fremur en þýðingu atriðanna í framhaldi af þessari rannsókn. Atriði 6, „*Þegar ég er í tilfinningalegu uppnámi hugsa ég ekki skýrt*“, reyndist einnig fremur flókið í þýðingu en niðurstöður viðtala bentu til þess að skilningur þátttakenda hafði verið svipaður og túlkun þýðenda á upprunalegri útgáfu atriðins, það er að „*tilfinningalegt uppnám*“ gæti átt við jákvætt og neikvætt tilfinningalegt ástand. Ekki var því talin ástæða til að breyta þýðingu þessa atriðis. Atriði 34, „*Ég les aldrei bækur eða tímarit með kynþokkafullu innihaldi*“, reyndist fremur erfitt í þýðingu. Þátttakendur túlkuðu þessa spurningu á ólíkan hátt og þá sérstaklega hvers kyns ritefni félli undir þessa skilgreiningu. Þátttakendur töldu þó samskonar bækur og tímarit falla

undir íslenska þýðingu atriðisins og upprunalega útgáfu atriðisins þannig að ekki er talið að vankantar séu á þýðingu atriðisins. Vinsældir bókar á þessu sviði er þó talin hafa áhrif á svör þátttakenda við þessu atriði þar sem flestir nefndu ákveðna bók sem dæmi um slíkt ritefni. Enn fremur er mögulegt að vinsældir þeirrar bókar hafi haft áhrif á lága fylgni atriðisins við önnur atriði þar sem þátttakendur voru opinskáir varðandi lestur bókarinnar og virtust ekki upplifa þá hegðun sem félagslega óæskilega.

Umræða

BIDR-6 kvarðinn á að mæla félagslega æskilega svörun FÆS og er yfirleitt notaður til að kanna hvort mælingar með öðrum sjálfsmatskvörðum séu bjagaðar vegna slíkrar svörunar. Ástæða er til að ætla að bjögun vegna FÆS sé vandamál þegar íslenskir sjálfsmatskvarðar eru notaðir ekki síður en erlendir. Því er þörf á tæki eins og BIDR-6 kvarðanum hér á landi, sem mælir FÆS á áreiðanlegan og réttmætan hátt. Markmið rannsóknarinnar var að þýða BIDR-6 úr ensku yfir á íslensku og kanna síðan próffræðilega eiginleika íslenskrar útgáfu kvarðans. Helstu niðurstöður benda til þess að þýðing mælitækisins hafi tekist nokkuð vel. Meðaltöl voru á svipuðu bili og erlendis. Ímyndarstjórnun mældist þó aðeins hærri hér á landi en erlendis (Holden o.fl, 2000; Li og Reb, 2009; Stöber o.fl., 2002). Nokkrar mögulegar ástæður eru taldar geta útskýrt mun á meðaltölum ÍS sem kanna þyrfti nánar. Í fyrsta lagi gæti þýðingu atriðanna sem mynda þáttinn verið ábótavant. Í öðru lagi gæti verið til staðar menningarmunur á skilningi atriðanna, tíðni hegðunar og hversu tilbúið fólk er til að viðurkenna hegðunina en rannsóknir benda til þess að menningarmunur geti verið í svörum við BIDR (Lalwani, Shavitt og Johnson, 2006; Lalwani, Shrum og Chiu, 2009). Í þriðja lagi gæti framkvæmd rannsóknar hafa haft áhrif og þeir þátttakendur sem notaðir voru í úrvinnslu einfaldlega frekar haft tilhneigingu til að reyna að ýkja jákvæða eiginleika í eigin fari en erlendir samanburðarhópar. Aðeins voru þeir þátttakendur notaðir í úrvinnslu gagna sem svöruðu öllum atriðum á listanum. Jákvæð fylgni hefur verið á milli samviskusemis á NEO-FFI persónuleikaprófinu og ÍS í fyrri rannsóknum (Stöber o.fl., 2002) og neikvæð fylgni milli samviskusemis og þess að sleppa þátttöku eða einstökum atriðum (Rogelberg, Conway, Sederburg, Spitzmüller, Aziz og Knight, 2003). Mögulegt að þeir þátttakendur séu hærri á ÍS en þeir sem ekki svöruðu öllum atriðum og gæti það

mögulega skýrt hærra meðaltal á íslenska kvarðanum. Í þeim erlendu úrtökum sem notuð hafa verið til samanburðar hefur kvarðinn oftast verið lagður fyrir í kennslustund þar sem svarhlutfall hefur verið mun hærra (t.d. Pauls og Crost, 2004 og Pauls og Stemmler, 2003). Innri áreiðanleiki beggja undirkvarða var hærri en 0,7 og bendir það til þess að áreiðanleiki íslenskrar útgáfu BIDR-6 sé ásættanlegur (DeVellis, 2003).

Niðurstöður staðfestandi þáttagreiningar styðja almennt séð réttmæti beggja undirkvarða listans. Fyrir báða undirkvarða féll eins þátta líkan nokkuð vel að gögnunum þó svo að fylgni á milli mælivillna nokkurra atriða drægi nokkuð úr mátgæðum. Flest atriðin virðast því mæla sama hugtak en þó ber að hafa í huga að viss atriði höfðu lága þáttahleðslu. Niðurstöður viðtala bentu til þess að vandamál við þýðingu gætu verið til staðar varðandi atriði 13, 16, 23, 25 og 39. Talið var að orðalag atriða 13, 23, 25 og 39 gæti verið of sterkt í íslenskri þýðingu. Af þessum atriðum er það helst atriði 13 sem kom jafnframt ekki nægilega vel út úr tölfræðilegri greiningu. Þegar svör fólks við því atriði voru skoðuð nánar bentu niðurstöður þó til þess að yfir helmingur svarenda valdi tvo sterkustu svarmöguleikana og voru því mjög sammála fullyrðingunni. Meðaltal þessa atriðis var með því hæsta í rannsókninni af öllum atriðum og studdi því tölfræðileg greining ekki að atriðið væri of sterkt orðað. Atriði 23, 25 og 39 komu einnig ágætlega út úr tölfræðilegri greiningu og bentu meðaltöl atriða ekki til þess að atriðin væru of sterkt orðuð. Mögulegt er að hægt sé að bæta þýðingu atriða 16, sem er líklega of flókið í íslenskri þýðingu en tölfræðileg greining benti þó til þess að atriðið væri ásættanlegt.

Niðurstöður staðfestandi þáttagreiningar sýndu einnig að tvíkosta skorun kemur betur út en samfelld í samræmi við niðurstöður Leite og Beretvas (2005). Samkvæmt Paulhus (1984) er ætlunin með BIDR að bera kennsl á ýkta svörun og benda niðurstöður þessarar rannsóknar til þess að tvíkosta skorun geti verið ákjósanlegri. Niðurstöður viðtala benda jafnframt til þess að fjöldi svarmöguleika sé ef til vill of mikill og fólki finnist erfitt að velja á milli sjö valmöguleika þegar kemur að því hversu sönn fullyrðing er. Innri áreiðanleikinn var þó örlítið hærri fyrir samfellda skorun í samræmi við niðurstöður Stöber og félaga (2002). Mikilvægt væri þó að framkvæma frekari rannsóknir til að kanna mun á tvíkosta og samfelldri skorun þar sem niðurstöður fyrri

rannsókna (Stöber o.fl., 2002) benda til þess að samfelld skorun sé ákjósanlegri með tilliti til áreiðanleika, samleitniréttmætis og fylgni við persónuleikapróf.

Fyrstu niðurstöður benda því til að íslensk útgáfa BIDR-6 kvarðans sé áreiðanleg og réttmæt og þá sér í lagi ÍS kvarðinn með tvíkosta skorun. Þó gæti verið nauðsynlegt að lagfæra þýðingu einhverra atriða. Rétt er að benda á að þessi rannsókn er einungis fyrsta skrefið í mati á próffræðilegum eiginleikum íslenskrar þýðingar BIDR-6 kvarðans. Því er mikilvægt að gera frekari rannsóknir á áreiðanleika og réttmæti BIDR-6 kvarðans áður en hægt er að mæla með almennri notkun hans. Það er til að mynda mikilvægt að kanna próffræðilega eiginleika hans í breiðara og stærra úrtaki. Einnig er nauðsynlegt að kanna réttmæti kvarðans með fjölbreyttari aðferðum eins og fylgni við aðra kvarða sem mæla kenningarlega tengd hugtök eða gagnsemi hans við að greina á milli þeirra sem svara á félagslega æskilegan hátt á öðrum mælingum. Eins er talið mikilvægt að framkvæma ítarlegri tölfræðigreiningar, svo sem svarferlalíkön (*item response theory*), þar sem kanna mætti nánar hvernig svarmöguleikarnir sjö nýtast. Einnig væri gagnlegt að fá frekari upplýsingar um skilning fjölbreyttari hóps á hugtökum í upprunalegri útgáfu kvarðans þrátt fyrir að tvítyngdir þýðendur hafi komið að þýðingu þessa mælitækis. Gagnlegt hefði verið að hafa aðgang að slíkum gögnum svo hægt væri að leggja betur mat á það hvort vandamál stafi af þýðingu eða innihaldi mælitækja. Mögulegt væri sem dæmi að leggja upprunalegu útgáfu mælitækisins fyrir nokkra viðmælendur sem annaðhvort hafa það tungumál sem móðurmál eða hafa dvalið lengi í landi þar sem tungumálið er talað. Að lokum gæti verið áhugavert að kanna hvort útbúa mætti styttri útgáfu af mælitækinu með því að fjarlægja atriði sem virðast eiga lítið sameiginlegt með öðrum atriðum. Dæmi eru um að styttar útgáfur af kvarðanum hafi verið notaðar erlendis af einstaka rannsakendum en þær útgáfur hafa verið þróaðar út frá einföldum aðferðum (Leite and Beretvas, 2005; Pauls and Stemmler, 2003; Stöber o.fl., 2002). Mælitækið er fremur langt og samanstendur af 40 atriðum og virðist því vera þörf á styttri útgáfu kvarðans. Niðurstöður viðtala benda enn fremur til þess að þrátt fyrir að ekki sé talið að vandamál sé til staðar varðandi þýðingu atriða séu ákveðin vandamál með upprunaleg atriði kvarðans, svo sem varðandi tvöfalda neitun. Mikilvægt er talið að kafa frekar ofan í þau atriði með tilliti til innihalds frekar en þýðingar.

Heimildir

Albright, J. J. og Park, H. M. (2009). Confirmatory factor analysis using Amos, LISREL, Mplus, and SAS/STAT CALIS. *Technical working paper.* Bloomington, IL: Indiana University.

Barrick, M. R. og Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44,* 1-26.

Barrick, M. R. og Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261-272.

Beatty, P. C. og Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287-311.

Behling, O. og Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. London: SAGE Publications.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246.

Beretvas, S. N., Meyers, J. L. og Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *62*(4), 570-589.

Cervellione, K. L., Lee, Y. S. og Bonanno, G. A. (2009). Rasch modeling of the self-deception scale of the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *69*(3), 438-458.

Costa, P. T. og MacCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI) professional Manual*. Odessa, FL: Psychological Assessment Resources.

Crowne, D. P. og Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349-354.

DeVellis, R. F. (2003). *Scale development: Theory and applications.* Thousand Oaks: Sage.

Einar Guðmundsson (2005-2006). Þýðing og staðfærsla sálfræðilegra prófa. *Sálfræðiritið – Tímarit Sálfræðingafélags Íslands, 10 – 11,* 23-40.

Flora, D. B. og Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks: Sage.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304-312.

Helmes, E. og Holden, R. R. (2003). The construct of social desirability: one or two dimensions? *Personality and Individual Differences*, *34*, 1015–1023.

Holden, R. R., Starzyk, K. B., McLeod, L. D. og Edwards, M. J. (2000). Comparisons among the Holden Psychological Screening Inventory (HPSI), the Brief Symptom Inventory (BSI), and the Balanced Inventory of Desirable Responding (BIDR). *Assessment*, *7*(2), 163-175.

Hu, L. T. og Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

Hu, L. T. og Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

International Test Commision (2000). ITC Guidelines on Adapting Tests. Sótt 11. desember 2013 af http://www.intestcom.org/upload/sitefiles/40.pdf

Jöreskog, K.G. (1993). Testing Structural Equation Models. Í Bollen, K.A. og Long, S.J. (ritstjórar), *Structural equation models* (bls. 295-316). California: SAGE Publications, Inc.

Jöreskog, K. G. og Sorbom, D. (2012). LISREL 9.1 [hugbúnaður]. *Lincolnwood, IL: Scientific Software International*.

Lalwani, A. K., Shavitt, S. og Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding?. *Journal of Personality and Social Psychology*, *90*(1), 165-178.

Lalwani, A. K., Shrum, L. J. og Chiu, C. Y. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology*, *96*(4), 870-882.

Lanyon, R. I. og Carle, A. C. (2007). Internal and external validity of scores on the Balanced Inventory of Desirable Responding and the Paulhus Deception Scales. *Educational and Psychological Measurement*, *67*(5), 859-876.

Leite, W. L. og Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne social desirability scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *65*(1), 140-154.

Li, A. og Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment*, *14*(2), 131-141.

Li, A. og Bagger, J. (2007). The Balanced Inventory of Desirable Responding (BIDR): A Reliability Generalization Study. *Educational and Psychological Measurement*, *67*(3), 525-544.

Li, A. og Reb, J. (2009). A cross-nations, cross-cultures, and cross-conditions analysis on the equivalence of the Balanced Inventory of Desirable Responding. *Journal of Cross-Cultural Psychology*, *40*(2), 214-233.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, *42(*5*)*, 851–858.

Mount, M. K., Barrick, M. R. og Stewart, G. L. (1998). Personality predictors of performance in jobs involving interaction with others. *Human Performance*, *11*(3), 145-166.

Ones, D. S. og Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*(2-3), 245-269.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias. Í J. P. Robinson, P. R. Shaver og L. S. Wrightsman (ritstjórar), *Measures of Personality and Social Psychological Attitudes* (bls. 17–59). San Diego: Academic Press.

Paulhus, D. L. (1994). Balanced inventory of desirable responding: Reference manual for BIDR version 6. Óbirt handrit, *University of British Columbia, Vancouver, Canada*.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. Í H. I. Braun, D. N. Jackson og D. E. Wiley (ritstjórar), *The role of constructs in psychological and educational measurement* (bls. 49-69). Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Pauls, C. A. og Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*(6), 1137-1151.

Pauls, C. A. og Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences*, *35*(2), 263-275.

Presser, S. og Blair, J. (1994). Survey pretesting: Do different methods produce different results. *Sociological Methodology*, *24*(1), 73-104.

Raykov, T. og Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling*, *13*(1), 130-141.

Rogelberg, S. G., Conway, J. M., Sederburg, M. E., Spitzmüller, C., Aziz, S. og Knight, W. E. (2003). Profiling active and passive nonrespondents to an organizational survey. *Journal of Applied Psychology*, *88*(6), 1104-1114.

Sackeim, H. A. og Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47*(1), 213-215.

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*,*76*(3), 323-346.

Sharma, S., Mukherjee, S., Kumar, A. og Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935-943.

Shevlin, M. og Miles, J. N. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, *25*(1), 85-90.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173-180.

Stöber, J., Dette, D. E. og Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment*, *78*(2), 370-389.

Vispoel, W. P. og Tao, S. (2013). A generalizability analysis of score consistency for the Balanced Inventory of Desirable Responding. *Psychological Assessment*, *25*(1), 94-104.

West, S., Taylor, A. og Wu, W. (2012). Model fit and model selection in structural equation modeling. Í R. H. Hoyle (ritstjóri), *Handbook of Structural Equation Modeling*. New York: Guilford.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Óbirt Doktorsritgerð, University of California, Los Angeles.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*(1), 115-148.

Zerbe, W. J. og Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12*(2), 250-264.

Short form development of the Balanced Inventory of Desirable Responding: Applying Confirmatory Factor Analysis, Item Response Theory, and Cognitive Interviews to Scale Reduction

Ragnhildur Lilja Asgeirsdottir, Fanney Thorsdottir, Vaka Vesteinsdottir

University of Iceland

Page title: Socially Desirable Responding: BIDR-6

Correspondence should be directed to Ragnhildur Lilja Asgeirsdottir, email: ragnhildurlilja@gmail.com.

Abstract

The Balanced Inventory of Desirable Responding (BIDR) is one of the most commonly used measures of socially desirable responding. It consists of two scales, Self-Deceptive Enhancement (SDE) and Impression Management (IM), containing 20 statements each, answered on a 7-point scale. The purpose of this paper was to analyze the items on the SDE and IM for a suggested short form version of the scale. Three studies were conducted. The first study (N=579) focused on analyzing the items using Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT), the second study (N=471) examined the effects of faking instructions on an item level and in the third study (N=20) cognitive interviews with probing were used in order to identify problematic items. Based on results from CFA, IRT, the effects of instructional variations and results of cognitive interviews, a short form version of the BIDR-6 was introduced, containing 10 SDE items and 10 IM items. The results suggest that the psychometric properties of measurements obtained with the two short forms were adequate.

Keywords: socially desirable responding, Balanced Inventory of Desirable Responding, short form development, confirmatory factor analysis, item response theory.

The purpose of this article is to develop a short form version of the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1991), one of the most commonly used measures of socially desirable responding. The BIDR is often used as a validation tool for other scales (e.g. Leite & Beretvas, 2005) even though the factorial stability of the scale has been criticized and items have been removed in previous studies to improve the fit of the model (Cervellione, Lee & Bonanno, 2009; Lanyon & Carle, 2007; Leite & Beretvas, 2005). The BIDR is also quite time-consuming, consisting of 40 items, which can be limiting, particularly when used in conjunction with other scales. Thus, a shorter, reliable and valid version of the BIDR would be beneficial.

Socially desirable responding

Personality questionnaires have been increasingly used in organizational settings. Although research has supported the use of such tools in personnel selection (Barrick & Mount, 1991; Mount, Barrick & Stewart, 1998; Salgado, 2003) other studies have indicated that applicants have a tendency to fake self-report measures by exaggerating their positive attributes and understating their negative attributes (Barrick & Mount, 1996; Ones & Viswesvaran, 1998; Zerbe & Paulhus, 1987), otherwise known as socially desirable responding (SDR). Paulhus has defined SDR as the tendency of respondents to give positive self-descriptions and classified it as a *response bias*, or a systematic tendency for responding on some other basis than item content, distinguishing between a *response style* (a bias consistent through time and situations) and *response sets* (a temporary bias due to situational demands) (1984; 1991; 2002). Different methods have been used to minimize the effect of SDR, e.g. forced-choice formats, randomized response method (see Greenberg, Abdula, Simmons & Horvitz, 1969 in Paulhus, 1991), asking a close acquaintance instead of the person in question, bogus pipelines (pseudo lie detectors; Jones & Sigall, 1971) and the most common method, anonymity. These methods focus on controlling for SDR, but in instances where SDR cannot be controlled or the aim is to directly measure SDR, the use of tests measuring SDR has become increasingly popular to identify or correct for SDR in self-report measures. Tests that measure SDR have mostly been used in three different ways: first,

by correlating the scores of the SDR measure and the scale of interest (low correlations indicating scores are not affected by SDR); second, by factor analyzing the scores of SDR measure and the focal scale (distinct factors indicating separate measures); and third, by deleting responses of those with high SDR scores (Beretvas, Meyers & Leite, 2002). Several SDR scales have been proposed (for a further reading on SDR scales, see e.g. Paulhus, 1991), the Marlowe-Crowne Social Desirability Scale (Marlowe & Crowne, 1960) and the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984) being the most commonly used SDR measures.

BIDR

According to Paulhus (1984; 2002) the BIDR was developed as an attempt to integrate the Self-Deception Questionnaire and the Other-Deception Questionnaire by Sackeim and Gur (1979) and the ideas of Damarin and Messick regarding an unconscious bias in self-reports on one hand and deliberate falsification on the other. The BIDR was developed by factor analyzing several SDR measures: the Edwards SD scale; the MMPI Lie scale; Wiggin´s SD scale; the Marlowe-Crowne scale; and the SDQ and ODQ scales (see Edwards, 1957; Meehl & Hathaway, 1946; Wiggins, 1959; Crowne & Marlowe, 1960; Sackeim & Gur, 1979, in Paulhus, 1984). The first four are scored in a true/false format and the latter two are answered on a 7-point scale but scored dichotomously (a point given to the two most extreme options). Although many of the items were rewritten, removed and added, the SDQ and ODQ items form the basis of the 40-item BIDR scale. The SDQ items loaded strongly on one factor, which Paulhus termed *self-deceptive enhancement* (SDE) and the ODQ items on another, *impression management* (IM). The SDE refers to the tendency to give self-reports that are honest but positively biased and the IM refers to the "deliberate self-presentation to an audience" (Paulhus, 1991, p.37). Although version 7, named the Paulhus Deception Scale (PDS), was released in 1998 for commercial use, version 6 (BIDR-6; Paulhus, 1991) continues to be the most commonly used scale in research and will be used in this study. The most significant change between the two versions is a change in the scoring system, from a 7-point scale in the BIDR-6 to a 5-point scale in the PDS and the two versions appear equivalent (Lanyon & Carle, 2007).

The BIDR-6 consists of 40 statements in total; 20 items representing SDE and 20 items representing IM. The items are answered on a 7-point scale, with 1 representing *not true*, 4 representing *somewhat true* and 7 representing *very true*. Items are alternated in content, with every other item scored reversely. Paulhus has allowed two different scoring methods, dichotomous and continuous[1] scoring (1994). With dichotomous scoring a point is given to each "6" and "7" response ("1" and "2" for reversed items), giving a maximum total of 20 points for each factor. With continuous scoring, negatively keyed items are reversed and points summed, giving a maximum score of 140 for each factor. Although Paulhus (1991) has recommended dichotomous scoring, ensuring that only those who exaggerate their responses may achieve high scores, previous studies have delivered conflicting results with regards to scoring methods. Research by Stöber, Dette and Musch (2002) has suggested that continuous scoring might be preferable with regards to reliability, convergent validity, effects of faking instructions and correlations with personality measures and findings from Rasch modeling (Cervellione et al., 2009) suggested collapsing the seven options into fewer options did not enhance the model fit. On the contrary confirmatory factor analysis results from Leite and Beretvas (2005) suggest dichotomous scoring may be preferable. Findings from research on the psychometric properties of measurements obtained with the BIDR scale have been mixed with regards to which scoring method may be preferrable. A meta-analysis of 110 studies by Li and Bagger (2007) suggested reliability (*Cronbach´s alpha*) of the BIDR was adequate, although the reliability coefficients ranged from .27 to .92 for SDE (mean=.68; standard deviation=.09) and .32 to .88 for IM (mean=.74; standard deviation=.09). The reliability was not statistically significantly related to the scoring method. Although Nunnally and Bernstein (1994) recommend reliability cutoffs of .80 for basic research, DeVellis (2003) has suggested alpha values between .65 and .70 to be minimally acceptable, values between .70 and .80 to be respectable, between .80 and .90 to be very good and values much above .90 suggesting redundant items. These guidelines apply to research instruments, higher standards are required for individual assessments that require critical accuracy,

[1] Strictly speaking, the measurements are discreet, thus *polytomous* would be a more correct descriptions, but *continuous* has traditionally been used to describe the seven point answering scale in the BIDR (e.g. Booth-Kewley, Edwards & Rosenfeld, 1992; Stöber et al., 2002).

suggesting that for research purposes, the reliability of the SDE scale is minimally acceptable and the IM scale is respectable. For higher stake situations, such as personnel selection, higher reliabilities would be needed. Test-retest reliability for SDE has been found to be .78 with dichotomous scoring, .83 for SDE with continuous scoring, .83 for IM with dichotomous scoring and .86 for IM with continuous scoring (Vispoel & Tao, 2013).

Although the psychometric properties of the BIDR-6 are acceptable, the factorial stability has been criticized and items have been removed in previous studies to improve the fit of the model (Cervellione et al., 2009; Lanyon & Carle, 2007; Leite & Beretvas, 2005). For this reason a shorter version of the BIDR-6 might be beneficial. In addition, the length of the BIDR-6 can limit its utility in research and industrial settings. Longer questionnaires are more time-consuming, result in more missing data and have higher refusal rates (Stanton, Sinar, Balzer & Smith, 2002).

Short form of the BIDR

Researchers have resorted to developing their own short forms of the BIDR, using methods such as confirmatory factor analysis (CFA) and corrected item-total correlations. An empirically based attempt by Leite and Beretvas (2005) to develop a short form version of the BIDR, using the standardized residuals in CFA, included 10 SDE items (3, 5, 6, 9, 11, 15, 16, 17, 19, and 20) and 10 IM items (24, 26, 29, 30, 31, 32, 34, 36, 37, and 38). Although the fit was not acceptable it appeared marginally close. Pauls and Stemmler (2003) reduced the scale to eight items on the SDE (1, 3, 7, 11, 12, 17, 18, and 19) and nine items on the IM (23, 26, 29, 30, 35, 36, 37, 39, and 40) by using Principal components analysis. Finally, Stöber, Dette and Musch (2002) used a short form consisting of 20 items, constructed based on factor loadings using exploratory factor analysis. The SDE consisted of items 1, 4, 5, 10, 12, 15, 16, 17, 18, and 20 and the IM contained items 21, 23, 24, 25, 29, 30, 33, 35, 36, and 37. No consensus has been reached by researchers regarding a short form of the BIDR and the main reason is probably the dependence on sample specific statistics in previous research. Another limitation of previous attempts is the emphasis on selecting items to maximize internal consistency. Focusing only on internal consistency in short form development may

create a short form that is too narrow and potentially low in validity (Loevinger, 1954). Thus, instead of selecting items for short form based only on internal consistency criteria, items should be selected on the basis of a number of qualities (Clark & Watson, 1995; DeVellis, 2003; Stanton, Sinar, Balzar & Smith, 2002). First, items should have internal qualities, which refer to measures of relations between items on the scale (e.g. means, corrected item-total correlations, factor loadings and item response theory). Second, items should have external qualities, which refer to the relationship between items on the scale and an external criteria and finally, judgmental qualities which refer to issues that require subjective judgment, such as clarity and understanding. A variety of these methods should be used for scale reduction, preferably retaining items with high scores on all three item qualities, giving preference to external item qualities, then internal item qualities and finally judgmental item qualities (Stanton, Sinar, Balzar & Smith, 2002).

The current research

The focus of this research was to propose a short form version of the BIDR-6 based on best practices in scale reduction (Clark & Watson, 1995; DeVellis, 2012; Stanton et al., 2002). Study 1 will focus on the internal item qualities with dichotomous and continuous scoring, using confirmatory factor analysis (CFA) and item response theory (IRT). Although previous research has analyzed the quality of items using CFA (Lanyon & Carle, 2007; Leite & Beretvas, 2005), this is the first study to analyze both SDE and IM with dichotomous and continuous scoring using IRT for this purpose. Study 2 will examine the effects of faking instructions on a scale and item level. Studies have indicated that both SDE and IM can be faked and scores are influenced by instructional variations (Paulhus, 2002; Paulhus, Bruce & Trapnell, 1995) and the same pattern will be expected for each item. It will be assumed that items that have low content validity, i.e. that people do not realize how to answer in a socially desirable way, are not valid measures of SDR, possibly due to complicated wording or behaviors that are no longer considered socially desirable. To the knowledge of the authors, this is the first attempt at examining the effect of faking instructions at the item level. Finally, Study 3 will

examine the judgmental item qualities, using cognitive interviews allowing for probing exploring qualities such as item clarity and perceived invasiveness of items.

Study 1

Method

Participants. The sample originally consisted of 649 participants. 403 of those were university students who volunteered to take part in the study and the remaining 246 participants were office workers recruited in an effort to make the sample more diverse. Since the aim of this study was to develop a short form version of the BIDR-6, it was decided to include only those participants who answered all questions on the BIDR-6, in total 579 participants. The mean age was 34.27 (SD=10.46; range 19-71; 6 participants did not indicate their age) and 19% were male and 81% were female (6 participants did not indicate their gender).

Measure and procedure. The BIDR-6 (Paulhus, 1994) contains 40 items, which consist of two 20 item subscales; Self-Deceptive Enhancement (SDE) and Impression Management (IM). All items are stated as propositions with a 7-point answer scale (1=*not true*, 4=*somewhat true* and 7=*very true*) and half of the items on each subscale are negatively keyed. In previous research means have ranged from 3.9 to 6.2 for SDE with dichotomous scoring, 77.2 to 87.9 for SDE with continuous scoring, 4.4 to 6.5 for IM with dichotomous scoring and 66.3 to 79.7 for IM with continuous scoring (Holden et al., 2000; Li & Reb, 2009; Stöber et al., 2002; Vispoel & Tao, 2012). A translated (Icelandic) version of the BIDR-6 was used in this study. Previous research on the Icelandic version suggests it is comparable to the original version (Asgeirsdottir, Thorsdottir & Vesteinsdottir, 2014). The questionnaire was computer administered; subjects received an email requesting participation in the study and were instructed to answer the questions to the best of their ability, with regards to how true each statement was.

Data analysis. Due to conflicting results from previous research (Cervellione et al., 2009; Stöber et al., 2002) and the recommendations by Paulhus (1991), both continuous and dichotomous scoring were examined. Continuous scoring on the BIDR-6

is computed by reverse scoring negatively keyed items and summing all answers, while dichotomous scoring is done by assigning a score of 1 point to each "6" and "7" answer and 0 points to other answers on positively keyed items and 1 point to each "1" and "2" answer and "0" to other answers on negatively keyed items. To identify problematic items confirmatory factor analysis (CFA) and item response theory (IRT) were used.

CFA was done using Lisrel 9.1 (Jöreskog and Sörbom, 2012). The method of estimation was diagonally weighted least squares (DWLS), due to a relatively small sample size and skewed distributions (*Shapiro-Wilk* tests were significant; Flora and Curran, 2004). When using DWLS the polychoric (tetrachoric for dichotomous variables) correlation matrix is analyzed. The Satorra-Bentler χ^2 (SBχ^2), Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI) and Non-normed Fit Index (NNFI), also known as the Tucker-Lewis Index (TLI), were used to measure the fit of the model, according to recommendations by Hu and Bentler (1998) and Yu (2002). The SBχ^2 (lower and non-significant chi-square suggests that the model fits the data) does have a tendency to over reject models at smaller sample sizes (Yu, 2002) but the SB-based RMSEA, CFI and NNFI are less sensitive to sample size (Bentler, 1990; Bollen, 1990; Hu & Bentler, 1995; 1998; 1999). The higher the RMSEA values are, the worse the model fits the data. Values $\leq .05$ or $\leq .06$ are considered to indicate a good fit, between .05 or .06 and .08 an adequate fit, and between .08 and .10 a mediocre fit (Browne & Cudeck, 1993; Hu & Bentler, 1999; Steiger, 1990). CFI and NNFI values over .95 or .96 are generally considered to indicate an excellent fit (Bentler, 1990; Hu & Bentler, 1999; Yu, 2002) and values over .90 an acceptable fit (Yuan, 2005). Factor loadings will also be used to identify problematic items using the guideline that factor loadings <.30 are considered low (e.g., Sharma, Mukherjee, Kumar og Dillon, 2005; Shevlin og Miles, 1998).

IRT has been a popular method for evaluating educational instruments but its use in evaluating responses to personality measures is increasing. In IRT the responses to a test are analyzed, assuming differing degrees of item difficulty and individual variation in trait levels. For scale reduction IRT can be a powerful tool, allowing for a more detailed analysis of responses and identifying items that effectively discriminate

between those who have higher levels of socially desirable responding and those who have lower levels, and for identifying and retaining items with a range of item difficulties. Rasch modeling, a type of IRT, has been used to analyze the properties of the SDE (Cervellione et al., 2009) but a detailed analysis of the item qualities of the BIDR using these methods has not been conducted for the purpose of reducing the scale length. The guidelines for applying IRT to scale reduction suggested by Edelen and Reeve (2007) will be followed in this study.

IRTPRO (Cai, L., Thissen, D. & du Toit, 2011) was used for IRT calculations. Underlying assumptions of IRT are the unidimensionality of the scale and local independence, meaning test items are pairwise uncorrelated if ability level (θ) is held constant. The assumptions will be estimated using CFA. For dichotomous scoring a two parameter logistic (2PL) model was used and for continuous scoring the graded response model was used (GRM). The item response function (IRF) describes the relation between the trait level and the probability of a certain item response. Theta, θ, represents the ability or trait level (in this case a function of the total SDE and IM scores). The *a* parameter equals the discrimination or the slope of the line. *a* values tend to range from 0.5 to 1.5, a higher *a* indicates the item discriminates better between those of lower ability and those of higher ability (Reise and Henson, 2003). According to Baker (2001), an approximate rule for interpreting *a* values states that values of 0 offer no discrimination, 0.01-0.34 very low discrimination, 0.35-0.64 low discrimination, 0.65-1.34 moderate discrimination, 1.35-1.69 high discrimination, and >1.70 very high discrimination. The *b* parameter represents the item difficulty, or the point of θ where the IRF has the greatest slope or the probability of endorsing an item is 0.5. *b* values close to zero imply medium difficulty compared to other items, items with negative values are generally considered easier and more frequently endorsed and positive values indicate items that are more difficult and less frequently endorsed (Reise and Henson, 2003). For continuous scoring, six *b* values are given indicating the six thresholds between the seven categories of the answer scale. The IRT is commonly used in the education field, but when applied to personality the ability applies to the amount a person exhibits of a certain trait. Thus difficult items, i.e. items with a higher *b*, imply that individuals have to show a higher level of socially desirable responding in

general (have a higher total score on the SDE and IM scales) to answer that question in a socially desirable way. The standard error of estimate (SEE) represents the uncertainty of a person´s location, the higher the SEE the greater the uncertainty. The inverse of the SEE is the item information which can be represented with the item information function (IIF). The IIF reflects the level of precision in item estimates at differing levels of ability (θ). The sum of the IIFs can be analyzed with the total information function (TIF), indicating the level of precision the test offers at different ability levels. For each item in this study the discrimination parameters (*a*), item difficulties (*b*) and test information functions, which indicate how effective an item is in measuring the trait at different trait levels, were analyzed. Preference will be given to items offering maximum information, particularly those discriminating well among those with higher levels of SDE and IM, since the purpose of the BIDR-6 is to identify those exhibiting high levels of socially desirable responding. Items offering little or redundant (i.e. similar item difficulty and discrimination) information will be recommended for deletion.

Results and discussion

Descriptive statistics. Means, standard deviations, reliability coefficients and intercorrelations of subscales, both from dichotomous and continuous scoring, are presented in Table 1 (N=579). The internal reliability of the SDE and IM was estimated with Cronbach´s alpha.

Table 1. Means, standard deviations, Cronbach´s alpha and intercorrelation.

					Correlation		
		M	*SD*	*α*	*1*	*2*	*3*
SDE	Dichotomous	4.93	3.10	.68			
SDE	Continuous	83.00	11.68	.69	0.76*		
IM	Dichotomous	6.67	3.69	.76	0.37*	0.35*	
IM	Continuous	80.80	16.00	.77	0.25*	0.39*	0.86*

Note: *p<0.01

The reliability of the subscales was slightly higher with continuous scoring. Means for the SDE and IM were similar to means found using the original version and the reliability was very close to the mean reliability found in a meta-analysis of reliability

coefficients in 110 studies on the BIDR-6 (Holden et al., 2000; Li & Bagger, 2007; Li & Reb, 2009; Vispoel & Tao, 2013). Means for IM were slightly higher than those found in previous studies, which can be a result of the sample used. Participation in the study was voluntary and only answers from those participants who answered every question were used. Previous results suggest a significant positive correlation between IM and conscientiousness (Stöber et al., 2002) and a negative correlation between conscientiousness and nonresponse (Rogelberg, Conway, Sederburg, Spitzmüller, Aziz & Knight, 2003) providing a possible explanation for slightly higher means. The correlation between subscales was also similar to correlations found in a previous study (Stöber et al., 2002) and highest between different scoring methods of the same scale.

Confirmatory factor analysis. CFA was used to examine the fit of the two models, SDE and IM. The two factors were analyzed separately, mostly due to the SDE and IM commonly being used separately. The variance of each factor was constrained to equal 1. The SBχ^2 fit indices suggest that none of the models fit the data as can be seen in Table 2. The RMSEA fit indices suggest an adequate fit for dichotomous SDE and IM scoring and a mediocre fit for continuous scoring for both factors. CFI and NNFI fit indices suggest a good fit for dichotomous SDE and IM scoring. The results of CFA thus suggest a better fit for dichotomous scoring than continuous scoring. The results are comparable to previous studies, albeit suggesting a slightly better fit (Leite & Beretvas, 2005).

Table 2. CFA results for dichotomous and continuous scoring for SDE and IM.

		SBχ^2	*df*	*p*	*RMSEA*	*Confidence interval for RMSEA (90%)*	*CFI*	*NNFI*
SDE	Dichotomous	357.43	170	0.00	0.072	0.066-0.078	0.963	0.958
	Continuous	515.21	170	0.00	0.083	0.077-0.089	0.848	0.830
IM	Dichotomous	357.59	170	0.00	0.060	0.054-0.066	0.975	0.972
	Continuous	558.80	170	0.00	0.081	0.076-0.087	0.901	0.890

Factor loadings for the SDE scale were analyzed in order to identify potential problematic items and the findings are presented in Table 3.

Table 3. Factor loadings (CFA) for SDE with dichotomous and continuous scoring.

	Factor loadings	
	Dichotomous	*Continuous*
1	***0.25***	***0.27***
2	*0.45*	*0.42*
3	***0.19***	***0.07****
4	*0.40*	***0.28***
5	*0.61*	*0.41*
6	*0.61*	*0.53*
7	***0.20***	***0.04****
8	***0.25***	***0.10****
9	***0.32***	***0.38***
10	*0.58*	*0.53*
11	*0.50*	*0.42*
12	*0.44*	***0.39***
13	***0.27***	***0.12***
14	***0.36***	***0.25***
15	*0.60*	*0.49*
16	*0.45*	***0.30***
17	*0.65*	*0.44*
18	*0.46*	*0.43*
19	*0.42*	***0.27***
20	*0.57*	*0.57*

Note: *p>0.05

As can be seen in Table 3, factor loadings for SDE ranged from .19 to .65 for dichotomous scoring. The factor loadings of five items were low (1, 3, 7, 8, and 13) with two items just above the .30 guideline, items 9 and 14. The factor loadings for SDE with continuous scoring ranged from .12 to .57, although three items (3, 7 and 8) did not load on the factor. Five other items fell below .30, items 1, 4, 13, 14, and 19, and three items (9, 12 and 16) were still quite low (<.40). The results again suggest dichotomous scoring is preferable for SDE. The factor loadings for the items on the IM scale are presented in Table 4.

Table 4. Factor loadings (CFA) for IM with dichotomous and continuous scoring.

	Factor loadings	
	Dichotomous	*Continuous*
21	*0.72*	*0.60*
22	*0.65*	*0.42*
23	*0.64*	*0.58*
24	***0.38***	*0.40*
25	*0.54*	*0.51*
26	*0.41*	*0.44*
27	*0.69*	*0.56*
28	*0.50*	***0.37***
29	***0.36***	***0.30***
30	*0.43*	***0.35***
31	***0.36***	***0.36***
32	***0.36***	***0.35***
33	*0.42*	*0.44*
34	***0.27***	***0.21***
35	***0.38***	***0.34***
36	*0.48*	***0.36***
37	*0.49*	*0.46*
38	*0.60*	*0.43*
39	*0.57*	*0.44*
40	*0.61*	*0.43*

Factor loadings are higher for IM than SDE as can be seen in Table 4. Factor loadings for IM ranged from .27 to .72 for dichotomous scoring and .21 to .60 for continuous scoring. Item 34 is the only item with a factor loading of less than .30 for both scoring methods. For dichotomous scoring the factor loadings of five items were just above the .30 guideline, items 24, 29, 31, 32, and 35 and for continuous scoring seven items (28, 29, 30, 31, 32, 35, and 36) were quite low (<.40).

Item response theory. Item parameters were first estimated for SDE with dichotomous and continuous scoring and are presented in Table 5.

Table 5. 2PL and GRM parameter estimates for SDE with dichotomous and continuous scoring.

	Dichotomous		*Continuous*					
Item	*a*	*b*	*a*	b_1	b_2	b_3	b_4	b_5
1	0.45	3.98	0.49	-7.87	-5.69	-3.54	0.98	3.64
2	0.88	1.91	0.85	-4.30	-2.28	-0.91	0.73	1.96
3	0.31	7.43	0.14	-15.46	-5.86	1.20	11.34	16.59
4	0.82	2.26	0.52	-4.75	-2.16	-0.41	1.93	3.25
5	1.30	1.12	0.76	-4.79	-2.97	-1.25	0.25	1.64
6	1.38	2.00	1.15	-2.19	-0.94	0.10	1.33	2.23
7	0.34	5.74	0.05	-49.38	-20.65	-0.09	17.94	35.88
8	0.45	2.78	0.20	-9.41	-4.99	-2.02	2.65	6.01
9	0.59	3.34	0.71	-3.02	-1.44	-0.40	1.33	2.85
10	1.24	1.31	1.16	-2.57	-1.45	-0.47	0.39	1.35
11	1.14	2.91	0.84	-2.07	-0.16	1.36	2.65	3.65
12	0.83	1.75	0.83	-3.77	-2.01	-0.59	0.77	1.75
13	0.45	-0.52	0.21	-18.61	-13.73	-10.19	-4.42	-1.10
14	0.62	0.50	0.45	-6.19	-4.23	-2.74	-0.68	0.67
15	1.20	0.91	0.92	-3.52	-2.43	-1.49	-0.14	1.07
16	0.85	0.89	0.60	-7.04	-5.10	-2.86	-1.01	1.17
17	1.39	0.01	0.84	-6.66	-5.31	-3.67	-1.81	0.04
18	0.85	1.27	0.94	-3.40	-1.86	-0.77	0.46	1.18
19	0.74	0.95	0.48	-7.90	-4.45	-2.58	-0.45	1.38
20	1.19	1.23	1.29	-2.67	-1.63	-0.69	0.49	1.14

Note: a = discrimination parameters, *b* = item difficulties, b_1, b_2, b_3, b_4 and b_5 =threshold parameters.

When analyzing the parameter estimates, Baker´s (2001) guidelines for interpreting item discrimination (*a*) were followed. As can be seen in Table 5, the *a* values for SDE with dichotomous scoring ranged from 0.31 to 1.39 (standard error (S.E.) range 0.11-0.25) and items 1, 3, 7, 8, 9, 13 and 14 had *a* values below 0.65, indicating discrimination is low or very low. The same items were identified as problematic by CFA, having factor loadings below .40. Table 5 also shows the parameter estimates for SDE with continuous scoring, with *a* values ranging from 0.05 to 1.29 (S.E. range 0.09-0.13), indicating that items 1, 3, 4, 7, 8, 13, 14, 16 and 19 had low or very low *a* values, offering little discrimination. These items also had low (≤.30) factor loadings in CFA. Certain *b* values (item/category difficulties) are quite extreme, most notably for items 3 and 7 with dichotomous scoring and items 1, 2, 3, 4, 5, 7, 8, 13, 14, 16, 17, and 19 with

continuous scoring, indicating problematic items. Item parameters were next estimated for IM with dichotomous and continuous scoring and findings are presented in Table 6.

Table 6. 2PL and GRM parameter estimates for IM with dichotomous and continuous scoring.

	Dichotomous		*Continuous*					
Item	*a*	*b*	*a*	b_1	b_2	b_3	b_4	b_5
21	1.73	0.71	1.38	-2.71	-1.61	-0.81	0.30	0.77
22	1.54	1.10	0.98	-3.42	-1.72	-0.23	0.51	1.38
23	1.39	0.68	1.32	-3.49	-2.11	-0.96	0.13	0.69
24	0.81	3.37	0.69	0.05	1.31	2.25	3.24	3.85
25	1.04	-0.51	1.07	-4.94	-3.22	-2.29	-1.18	-0.51
26	0.72	-0.04	0.88	-3.96	-2.58	-1.52	-0.81	-0.08
27	1.60	1.27	1.18	-2.12	-0.91	-0.02	1.16	1.53
28	0.97	0.86	0.76	-4.25	-2.47	-1.35	0.29	1.00
29	0.62	0.32	0.57	-3.58	-2.15	-1.23	-0.19	0.37
30	0.80	1.95	0.64	-1.40	-0.15	0.80	1.76	2.31
31	0.64	-0.12	0.65	-3.53	-2.48	-1.43	-0.48	-0.10
32	0.63	1.80	0.58	-1.72	-0.19	0.79	1.50	1.95
33	0.85	2.73	0.82	-1.23	0.11	1.33	2.36	2.83
34	0.49	3.82	0.35	-2.32	0.80	2.36	4.10	5.21
35	0.80	2.83	0.63	-0.84	0.64	1.74	2.88	3.43
36	0.92	-0.53	0.73	-4.10	-2.92	-1.98	-1.14	-0.66
37	0.93	0.34	0.91	-1.86	-1.06	-0.48	0.11	0.36
38	1.30	-0.90	0.83	-3.33	-2.48	-2.09	-1.56	-1.18
39	1.20	1.11	0.88	-2.57	-1.34	-0.51	0.74	1.33
40	1.24	1.03	0.86	-3.70	-1.87	-0.68	0.52	1.29

Note: a = discrimination parameters, *b* = item difficulties, b_1, b_2, b_3, b_4 and b_5 =threshold parameters.

As shown in Table 6, *a* values for IM with dichotomous scoring range from 0.49 to 1.73 (S.E. range 0.11-.22), with items 29, 31, 32 and 34 having low or very low *a* values. In Table 6 it can also be seen that *a* values for IM with continuous scoring range from 0.35 to 1.38 (S.E. range 0.09-0.12), with items 29, 30, 32, 34 and 35 having low or very low *a* values. There was quite an agreement between IRT and CFA results. Item 34 was the only item that also had low factor loadings (<.30) in CFA, both for dichotomous and continuous scoring, but all other items identified as problematic in IRT also had rather low factor loadings (<.40). Although the *b* values (item/category difficulties) were not as extreme as for SDE, certain items were identified as problematic, most notably 25,

28 and 36 for continuous scoring. Test information functions for SDE and IM with dichotomous and continuous scoring are presented in Figure 1.

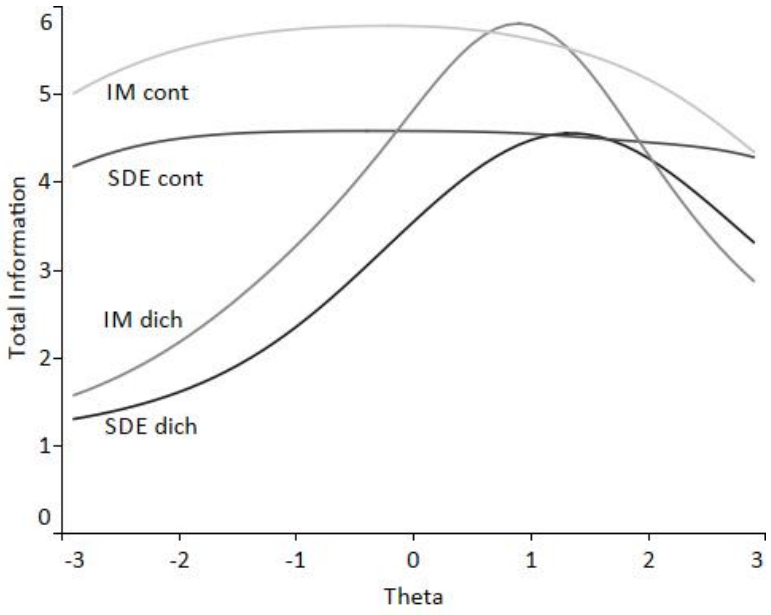

Figure 1. Test information functions for SDE and IM with dichotomous and continuous scoring.

Figure 1. shows the test information functions (TIF) for SDE and IM, both with dichotomous and continuous scoring. The TIFs for SDE and IM indicate that continuous scoring offers greater information across differing trait levels, while dichotomous scoring offers mostly information on those that exhibit more of the trait. Although, for SDE dichotomous scoring offers greater information than continuous scoring at θ (theta) levels of 1.2 to 1.6 and for IM dichotomous scoring offers greater information than continuous scoring at θ levels of .7 to 1.2. The TIFs indicate that both scales would benefit from more items discriminating better among those at the highest trait levels, since the aim of the BIDR-6 is to identify those who fake their answers to appear as socially desirable as possible. The item information functions were also analyzed (although not reported here due to length constraints) for all items on the SDE and IM with dichotomous and continuous scoring, taking into consideration the TIFs. Although

most of the items offering little information had already been identified through analysis of item discrimination and item difficulties, four further items were suggested for item deletion, items 24 and 26 for IM with dichotomous scoring and items 24, 31 and 33 for IM with continuous scoring.

Study 2

Study 2 will examine the effects of faking instructions on a scale and item level. Studies have indicated that both SDE and IM can be faked and scores are influenced by instructional variations (Paulhus, 2002; Paulhus et al., 1995) although the IM does appear to be more sensitive to faking instructions (Holden, Starzyk, McLeod & Edwards, 2000; Paulhus et al., 1995; Reid-Seiser & Fritzsche, 2001; Stöber et al., 2002). Thus, it will be assumed that since the SDE and IM can be faked, the same pattern will be expected for each item. It will be assumed that items with low content validity, i.e. that people do not realize how to answer in a socially desirable way, are not valid measures of SDR. It will be hypothesized that items that are not influenced by instructional variations may be problematic, e.g. due to confusing wording or behaviors no longer considered socially desirable.

Method

Participants. The sample consisted of 552 university students and of those 471 completed the measure. Participants were randomly divided into two groups and received two different sets of instructions; standard instructions and faking instructions. Participants completing the measure under normal instructions were 258 (22.5% males, 77.5% females). Mean age was 32.67 (SD=10.29, range 19-71). 213 participants completed the measure under faking instructions (18.8% males, 80.8% females). Their mean age was 32.14 (SD=10.32, range 19-60, 3 did not indicate gender).

Measure and procedure. As in study 1 the Icelandic translations of the BIDR-6 was used (Asgeirsdottir et al., 2014; Paulhus, 1994). Participants were recruited by email and the questionnaire was computer administered. The respondents were randomly divided into two groups: those who received standard instructions and those who received faking instructions. The faking instructions requested participants to fake their

responses in order to present themselves as favorably as possible according to social norms, and different scenarios were described in which people tend to do so, e.g., when applying for their dream job. Respondents were thus asked to respond as socially desirable as possible.

Data analysis. The effects of faking on scale and item means were evaluated. Independent samples T-tests were performed to test for significant differences between item means under standard instructions vs. faking instructions.

Results

Effects of faking on scale statistics. The effects of the instructions on means, standard deviations and internal reliability (*alpha*) can be seen in Table 7.

Table 7. Means, standard deviations and Cronbach´s alpha under standard and faking instructions

		Means		*Standard deviation*		*Alpha*	
		S	*F*	*S*	*F*	*S*	*F*
SDE	Dichotomous	4.75	9.99	2.91	5.21	.65	.87
SDE	Continuous	82.27	98.43	11.15	18.40	.66	.90
IM	Dichotomous	6.47	11.21	3.60	5.69	.75	.91
IM	Continuous	79.86	100.13	15.94	24.64	.77	.93

S=Standard instructions, F=Faking instructions

The results under standard instructions were very similar to the results from the sample described in study 1, although the means, standard deviations and internal reliability were slightly lower in this sample. As can be seen in Table 7 the SDE and IM means were considerably higher under faking conditions, both with dichotomous and continuous scoring. The difference between scores under standard and faking instructions were as follows: 5.24 ($t(469)=-13.758$ ($p<0.01$); SDE dichotomous scores); 16.16 ($t(469)=-11.739$ ($p<0.01$); SDE continuous scores); 4,74 ($t(469)=-10.983$ ($p<0.01$); IM dichotomous scores); and 20.27 ($t(469)=-10.764$ ($p<0.01$); IM continuous). Results for IM dichotomous under standard and faking conditions were almost identical to results from Holden (2007), 6.33 and 11.98 respectively. The faking instruction also

affected the reliability estimate, resulting in much higher values. Intercorrelations between factors under standard and faking instructions are presented in Table 8.

Table 8. Intercorrelations between factors under standard and faking instructions

		Correlation standard instructions			*Correlation faking instructions*		
		1	*2*	*3*	*1*	*2*	*3*
SDE	Dichotomous						
SDE	Continuous	0.74*			0.93*		
IM	Dichotomous	0.44*	0.40*		0.74*	0.72*	
IM	Continuous	0.28*	0.42*	0.85*	0.68*	0.70*	0.96*

Note: *p<0.01

The correlation between factors for the sample receiving standard instructions was very similar to the correlations described in study 1. Correlations between the BIDR-6 scales were much higher under faking conditions, in accordance with previous research, theoretically explained by a difference in the ability of individuals to fake (Pauls & Crost, 2004). These results are a good example of how correlations and reliability are biased upward when socially desirable responding affects measures.

Effects of faking on item means. The effect of faking on an item level with dichotomous scoring can be seen in Figure 2 and with continuous scoring in Figure 3.

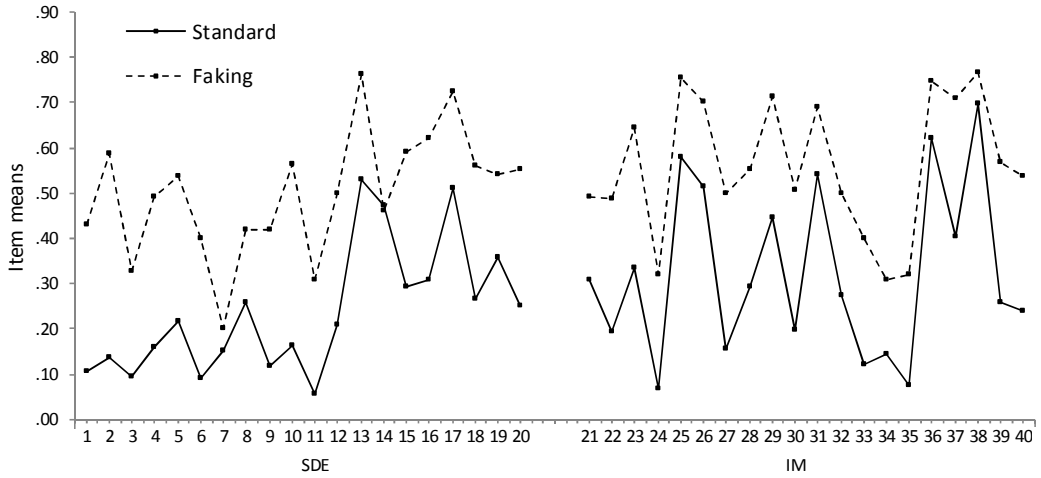

Figure 2. Dichotomous SDE and IM scores under standard and faking instructions

Significant effects of instructions were found for all SDE items, except item 7 (t(469)=-1.445 (p=.149)) and item 14 (t(469)=.276 (p=.783)) with dichotomous scoring and item 7 (t(469)=.842 (p=.400), item 8 (t(469)=-1.313 (p=.190)) and item 14 (t(469)=.449 (p=.654) with continuous scoring. For IM items, significant effects were found for all items, except item 38, both with dichotomous scoring (t(469)=-1.642 (p=.101)) and continuous scoring (t(469)=-1.134 (p=.257). Items not affected by faking instructions are interpreted as being poor measures of SDR.

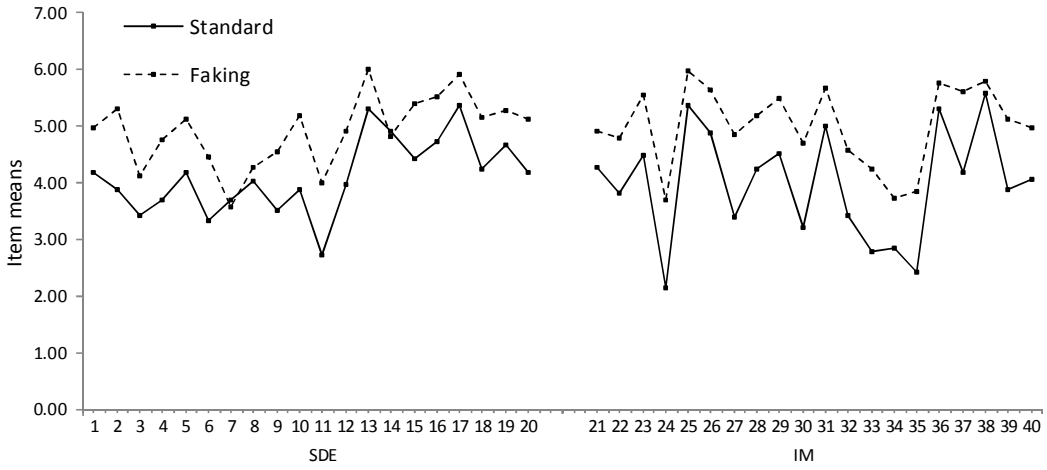

Figure 3. Continuous SDE and IM scores under standard and faking instructions

Study 3

Study 3 will examine the judgmental item qualities of the BIDR-6, using cognitive interviews allowing for probing, exploring qualities which refer to issues that require subjective judgment, such as clarity and understanding.

Method

Participants. Participants were 20; 10 males and 10 females. The choice of participants was made based on having a diverse sample with relation to gender, age and education. The age range was 21 to 58.

Measure and procedure. The participants were interviewed face-to-face in a quiet setting, most often in their own homes. Due to the length of the questionnaire the interviews were conducted in two settings, no more than two weeks apart. Each

interview focused on one scale and was on average 2 hours long (4 hours in total for each participant) and all interviews were conducted by the same interviewer (an undergraduate student in psychology trained in the method). The interviews were designed to explore the cognitive processes of participants while answering the items on the BIDR-6 (Icelandic translation; Asgeirsdottir et al., 2014) and were partly standardized, allowing for probing. The focus was mostly on the respondents' understanding of the questions and the applicability of the 7-point scale for answering each question. The respondents started by answering the BIDR-6 and were then asked to describe their thought process while arriving at their answers. The interviewer started with more general probing, followed by more specific probes, using guidelines from Fowler (1995) and Willis (2005). Four types of probes can be identified (Beatty & Willis, 2007): Anticipated probes (constructed prior to the interview (standardized) and initiated by the interviewer (proactive)); Emergent probes (constructed during the interview (non-standardized) and triggered by the respondent (reactive)); Conditional probes (standardized and reactive); Spontaneous probes (non-standardized and proactive). In this study a mixture of the four probes was used. Four specialists in the area prepared a script for the interviewer, which started off with open questions about the respondents' understanding of the item and how they arrived to their conclusion in choosing their answer, followed by specific questions about items, e.g., understanding of specific words or parts of sentences, and ending with an open question about the clarity of the item, allowing for probing initiated during the interview. Clarification of items was not allowed, thus questions from respondents were generally answered with questions asking the respondent what they thought the meaning was. The interviews were tape-recorded and transcribed. The transcripts were analyzed, items analyzed and key themes identified (Mallinson, 2002; Ritchie, Spencer & O'Connor, 2003). Participants were informed that they did not have to answer every question and did not have to explain or go into too many details about items or their answers if they did not feel comfortable doing so. Particular attention was made not to probe for further answers in items 18 and 34, since ethics committees have previously found those items to be too sensitive (Kam, 2013).

Results and discussion

Certain themes were identified when analyzing the transcripts concerning the comprehension of items, most specifically problems with double negations, difficult or vague questions, offensive questions and double barreled questions. Finally, the appropriateness of the answer scale also caused some difficulties with regards to certain items.

Double negation. Items that were identified as possibly being problematic due to double negations are items 3, 8, 11, 14, 20, 22, 24, 32, 34, 36, 37, 38 and 40. Special consideration should be given to item 3 ("*I don´t care to know….*") since 7 out of 20 subjects in this study originally answered the question in the opposite way they intended to.

Difficult or vague questions. Complicated or vague questions can easily present a problem to respondents. This can create a subtle problem if the difference in comprehension is slight but if the respondents' comprehension of the items differ greatly the effect can be quite detrimental. Items which were identified as possibly posing problems due to difficult or vague questions were items 3, 5, 7, 9, 13, 14, 16, 20, 21, 23, 25, 26, 27, 29, 30, 31, 34, 35, 36, 37, 39 and 40. An example of this is "*I am fully in control of my own fate*". Respondents interpreted the "*fully in control*" in different ways, some quite literally and others more leniently.

Offensive questions. In Study I comments were invited at the end of the survey. Two participants gave comments stating they found item 18 invasive and were offended being asked that question. With relation to other items on the scale they could not see how a question regarding their ability as a lover was relevant. Many participants in Study 3 chose not to answer this question in the face-to-face interview. Due to the diverse use of the BIDR-6 in studies it is recommended that item 18 should not be included. There have also been examples of ethics committees not allowing items 18 and 34 to be included in research (Kam, 2013), providing further strength to the argument. Items 34 and 35 were also considered too offensive for some respondents in this study to want to answer those items. It was noted though that the popularity of a

certain book falling under the category of a “sexy book” according to many respondents, at the time of the interviews, seemed to make it less socially undesirable to admit to reading books of that genre in response to item 34. Some respondents still found the item invasive.

Double barreled questions. One of the most important aspects of questionnaire designing is avoiding items which contain two questions, yet allowing only for one answer, i.e. double barreled questions. Items 13, 37 and 38 were identified as possibly being problematic due to this issue. Items 37 and 38 ask respondents to give a single answer to two different scenarios, for example regarding taking a sick-leave from school or work, which many respondents pointed out they would answer differently. Item 13, although not directly double barreled makes certain assumptions regarding voting behavior which some respondents had difficulty with (a one-and-a-half-barreled question; see Sudman and Bradburn, 1982 in Fowler, 1995).

Appropriateness of answer scale. The appropriateness of the answering scale was thought to be a problem for several items by many respondents. The most common reason for that was that the item contained information about frequency, e.g. “sometimes” and “usually”. Items 1, 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 21, 24, 25, 26, 27, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40 were all identified by some respondents as being difficult to answer according to the 7-point scale, though to differing degrees. In some instances the respondents thought it easier to answer the item on a true/false basis and in other instances they suggested the problem could be resolved by removing the word suggesting frequency from the item itself.

Suggested item revision

A summary of the results of studies 1, 2, and 3 for SDE can be seen in Table 9 and for IM in Table 10. Preference was given to internal item quality, i.e., low factor loadings, low discriminative powers and items not affected by faking instructions when reducing the length of the scale. In most cases items did not perform poorly on all of the above. In those cases subjective reasoning was used based on the results of Study 3. It was also decided to remove items 18, 34 and 35 due to their offensive content. The ten

items that performed the worst in the three studies were deleted from each factor. This number of items was chosen for convenience, resulting in two ten item scales.

Table 9. Summary of results from studies 1 (CFA and IRT), 2 (instructional variations) and 3 (interviews) for SDE

	Low factor loadings[a]		*IRT*[b]		*Faking t-test*[d]		*Item quality*[e]	*Item removal*
	D	*C*	*D*	*C*	*D*	*C*		
1	X	X	X	X			A	X
2				X			A	
3	X	X	X	X			NVA	X
4		X		X			A	X
5				X			VA	
6								
7	X	X	X	X	X	X	VA	X
8	X	X	X	X		X	NA	X
9			X				VA	
10								
11							NA	
12								
13	X	X	X	X			VBA	X
14		X	X	X	X	X	NV	X
15								
16				X			VA	X
17				X				
18							O	X
19		X		X				X
20							NV	

Note: D=Dichotomous, C=Continuous, [a] Factor loadings <.30, [b] Low item discrimination (<.64) or offers little information, [d] p>0.05 and [e] N=Double negation, V=Difficult or vague questions, O=Offensive questions, B=Double barreled questions, A=Appropriateness of answer scale.

Originally, nine items were selected for deletion (1, 3, 4, 7, 8, 13, 14, 18 and 19). For the tenth item selected for deletion the choice was between items 2, 5, 9 and 16. After reviewing the items, item 16 was selected for deletion, most notably due to redundant information provided by the item with dichotomous scoring, according to IRT results, as the item provided less information than items 5, 10, 15 and 20 at similar ability levels. IRT results had already indicated the item to be problematic with continuous scoring and lacking in clarity and appropriateness of answer scale. The suggested short form of the BIDR-6 thus consists of items 2, 5, 6, 9, 10, 11, 12, 15, 17 and 20 on the SDE scale. Five of the SDE items are negatively keyed, thus equaling the original proportion of

reversed items. A summary of the results of studies 1, 2 and 3 for IM are presented in Table 10.

Table 10. Summary of results from studies 1 (CFA and IRT), 2 (instructional variations) and 3 (interviews) for IM

	Low factor loadings[a]		*IRT*[b]		*Faking t-test*[d]		*Item quality*[e]	*Item removal*
	D	*C*	*D*	*C*	*D*	*C*		
21							VA	
22							N	
23							V	
24			X	X			NA	X
25				X			VA	
26			X				VA	X
27							VA	
28				X				
29			X	X			VA	X
30				X			VA	X
31			X	X			VA	X
32			X	X			NA	X
33				X			A	X
34	X	X	X	X			NVO	X
35				X			VOA	X
36				X			NVA	
37							NVBA	
38					X	X	NBA	X
39							VA	
40							NVA	

Note: D=Dichotomous, C=Continuous, [a] Factor loadings <.30, [b] Low item discrimination (<.64) or offers little information, [d] p>0.05 and [e] N=Double negation, V=Difficult or vague questions, O=Offensive questions, B=Double barreled questions, A=Appropriateness of answer scale.

The items on the IM performed better than the items on the SDE, with regards to CFA and faking instructions. Thus, the suggestions for the short form version of the IM scale depended mostly on IRT results. Items 34 and 35 were immediately suggested for deletion, due to their offensive content, and item 34 also having low factor loadings. Item 38 was recommended for deletion based on the effects of faking instructions. Items 24, 29, 31 and 32 were also recommended for deletion based on low discrimination and/or little information provided with both dichotomous and continuous scoring. After analyzing item discrimination and item information functions, items 26, 30 and 33 were also recommended for deletion. Thus, the suggested short form of the IM scale consists of items 21, 22, 23, 25, 27, 28, 36, 37, 39 and 40. Six of the

IM items are negatively keyed, thus almost retaining the original proportion of reversed items. The means, standard deviations, internal reliability (*alpha*) and correlations between factors for the short form suggestions for SDE and IM, with dichotomous and continuous scoring (based on sample in Study 1; N=579) can be seen in Table 11.

Table 11. BIDR-6 short form: means, standard deviations, Cronbach´s alpha and intercorrelations

					Correlation		
		M	*SD*	*α*	*1*	*2*	*3*
SDE	Dichotomous	2.21	1.92	.63			
SDE	Continuous	40.26	7.87	.70	.79*		
IM	Dichotomous	3.58	2.45	.72	.33*	.35*	
IM	Continuous	43.90	9.44	.72	.30*	.43*	.88*

Note: *p<0.01

The intercorrelations between factors are very similar to results from studies 1 and 2. The internal reliability (*alpha*) of the short form SDE and IM is slightly lower than for the full length version (from .01 to .05 lower), as can be expected since test length influences reliability (Li & Bagger, 2007; Nunnally & Bernstein, 1994). The alpha values for IM are the same with dichotomous and continuous scoring, but lower for SDE with dichotomous scoring than with continuous scoring. CFA results for the short form can be seen in Table 12. Although the results from the SBχ^2 suggest the models do not fit the data, the RMSEA, CFI and NNFI values indicate a good fit for SDE and IM with dichotomous scoring and an adequate fit for SDE and IM with continuous scoring.

Table 12. BIDR-6 short form: CFA results for SDE and IM with dichotomous and continuous scoring

		SBχ^2	*df*	*p*	*RMSEA*	*Confidence interval for RMSEA (90%)*	*CFI*	*NNFI*
SDE	Dichotomous	72.22	35	0.00	0.050	0.037-0.064	0.984	0.979
	Continuous	156.47	35	0.00	0.074	0.061-0.086	0.903	0.875
IM	Dichotomous	96.75	35	0.00	0.050	0.037-0.064	0.982	0.976
	Continuous	118.75	35	0.00	0.069	0.057-0.082	0.948	0.933

General discussion

The purpose of this study was to analyze the items on the BIDR-6 in order to provide a suggested short form version of the scale. Three studies were presented focusing on the quality of the items on the SDE and IM. The results of the first study suggested certain items had little in common with other items (CFA), as well as offering little discrimination between those with higher scores on the two scales and those with lower scores (IRT). Results from the second study showed that although both SDE and IM total scores increased under faking instructions, some items were not affected by instructional variations. The third study provided further information on the items through cognitive interviews with probing. Certain themes were identified and a majority of the items were identified as being problematic in some ways with regards to comprehension of the items or the appropriateness of the answer scale. Based on the results from studies 1, 2 and 3 a suggested short form version of the BIDR-6 was developed, containing 20 items. The reliability of the short form version was adequate and results from CFA indicated the scales were unidimensional. The SDE scale has eight items in common with the short form developed by Leite and Beretvas (2005) and the IM five items in common with the short form developed by Pauls and Stemmler (2003), suggesting certain items tend to perform better across studies.

Some potential limitations may be identified with the studies which must be taken into consideration. First, the CFA results and descriptive statistics for the short form version were calculated from the sample previously used in Study 1. Before further recommendations can be made regarding the use of the short form it would have to be validated using a separate sample in conjunction with other measures to establish validity and reliability, preferably a more diverse random sample. Second, a translated version of the BIDR-6 was used in this study, possibly affecting the results. Means, Cronbach´s alpha, intercorrelations and CFA results were similar to those found in previous research using the original version of the scale (Holden et al., 2000; Li & Bagger, 2007; Li & Reb, 2009; Vispoel & Tao, 2012), suggesting the Icelandic translation is adequate. Furthermore, the translation process was quite detailed, involving experts in the field and bilingual specialists, pretesting, and with previous results also

suggesting that the translation is adequate (Asgeirsdottir et al., 2014). Third, the majority of participants in the first two studies were females (from 77.5% to 80.8%). This can partly be explained by females accounting for approximately 65% of the students enrolled at the university at the time of the study (University of Iceland, 2013). Previous studies have also found a gender bias in web surveys, indicating women have lower nonresponse rates (Sax, Gilmartin & Bryant, 2003) although other studies have contradictory findings (Smith & Leigh, 1997).

Although, further research is needed on the short form version of the BIDR-6 presented in this article, before its use can be recommended, valuable information can be drawn from these results. First, the article provides a more detailed analysis of the items on the BIDR-6 than previously reported. It was decided to show results using both dichotomous and continuous scoring, due to conflicting recommendations regarding which is optimal (Cervellione et al., 2009; Paulhus, 1991; Stöber et al., 2002). This article does not provide a clearer suggestion on which method is preferable. Second, this paper provides an example of how a variety of methods can be used for scale reduction, most notably IRT. IRT is still quite a novel method for analyzing psychological tests, but the information it provides can be very beneficial. To the authors' knowledge this is the first attempt at using IRT for this purpose on the BIDR. Third, the methods used in this paper resulted in a suggested short form version of the BIDR-6, which if supported by further research, can be a valuable addition to the social desirability scales, being less time-consuming, adding to its utility.

References

Asgeirsdottir, R. L., Thorsdottir, F., & Vesteinsdottir, V. (2014). Félagslega æskileg svörun: Íslensk þýðing og próffræðilegir eiginleikar Balanced Inventory of Desirable Responding [Socially desirable responding: The psychometric properties of the Icelandic version of the BIDR]. Unpublished manuscript, University of Iceland, Iceland.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, *81*, 261-272.

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, *83*(3), 377-391.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287-311.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *62*(4), 570-589.

Bollen, K. A. (1990). Overall fit in covariance structure models: two types of sample size effects. *Psychological Bulletin*, *107*(2), 256-259.

Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77*(4), 562-566.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Ed.), *Testing structural equation models* (pp. 136-136). Newbury Park, CA: Sage Publications, Inc.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. *Chicago, IL: Scientific Software International*.

Cervellione, K. L., Lee, Y. S., & Bonanno, G. A. (2009). Rasch modeling of the self-deception scale of the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *69*(3), 438-458.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309-319.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349-354.

Edelen, M. O., & Reeve B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16* (1), 5-18.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216-1229.

Holden, R. R., Starzyk, K. B., McLeod, L. D., & Edwards, M. J. (2000). Comparisons among the Holden Psychological Screening Inventory (HPSI), the Brief Symptom Inventory (BSI), and the Balanced Inventory of Desirable Responding (BIDR). *Assessment*, *7*(2), 163-175.

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science, 39*(3), 184-201.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 76-99). Newbury Park, CA: Sage.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: a new paradigm for measuring affect and attitude. Psychological Bulletin, 76(5), 349-364.

Jöreskog, K. G., & Sorbom, D. (2012). LISREL 9.1 [Computer software]. *Lincolnwood, IL: Scientific Software International*.

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*(2), 201-228.

Kam, C. (2013). Probing item social desirability by correlating personality items with Balanced Inventory of Desirable Responding (BIDR): A validity examination. *Personality and Individual Differences,54,* 513-518.

Lanyon, R. I., & Carle, A. C. (2007). Internal and external validity of scores on the Balanced Inventory of Desirable Responding and the Paulhus Deception Scales. *Educational and Psychological Measurement*, *67*(5), 859-876.

Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne social desirability scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *65*(1), 140-154.

Li, A., & Bagger, J. (2007). The Balanced Inventory of Desirable Responding (BIDR): A Reliability Generalization Study. *Educational and Psychological Measurement*, *67*(3), 525-544.

Li, A., & Reb, J. (2009). A cross-nations, cross-cultures, and cross-conditions analysis on the equivalence of the Balanced Inventory of Desirable Responding. *Journal of Cross-Cultural Psychology*, *40*(2), 214-233.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*(5), 493.

Mallinson, S. (2002). Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. *Social Science & Medicine*, *54*(1), 11-21.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Personality predictors of performance in jobs involving interaction with others. *Human Performance*, *11*(3), 145-166.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*(2-3), 245-269.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L. (1994). Balanced inventory of desirable responding: Reference manual for BIDR version 6. *Unpublished manuscript, University of British Columbia, Vancouver, Canada*.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological*

and educational measurement (pp. 49-69). Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, *21*(2), 100-108.

Paulhus, D. L., & John, O. P. (1998). Egoistic and Moralistic Biases in Self-Perception: The Interplay of Self-Deceptive Styles with Basic Traits and Motives. *Journal of Personality*, *66*(6), 1025-1060.

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*(6), 1137-1151.

Pauls, C. A., & Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences*, *35*(2), 263-275.

Reid-Seiser, H. L., & Fritzsche, B. A. (2001). The usefulness of the NEO PI-R Positive Presentation Management Scale for detecting response distortion in employment contexts. *Personality and Individual Differences*, *31*(4), 639-650.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*(2), 93-103.

Ritchie, J., Spencer, L., & O'Connor, W. (2003). Carrying out qualitative analysis. *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, 219-262.

Rogelberg, S. G., Conway, J. M., Sederburg, M. E., Spitzmüller, C., Aziz, S., & Knight, W. E. (2003). Profiling active and passive nonrespondents to an organizational survey. *Journal of Applied Psychology*, *88*(6), 1104-1114.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47*(1), 213-215.

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*, *76*(3), 323-346.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, *44*(4), 409-432.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935-943.

Shevlin, M., & Miles, J. N. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, *25*(1), 85-90.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 496-505.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*(1), 167-194.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173-180.

Stöber, J., Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment*, *78*(2), 370-389.

University of Iceland. (2013). *University of Iceland in numbers* [*Hí í tölum*]. Retrieved 14. December, 2013, from https://docs.google.com/viewer?url=http%3A%2F%2F www.hi.is %2Fsites%2Fdefault%2Ffiles%2F skradir_feb_2013_heild.xlsx

Vispoel, W. P., & Tao, S. (2013). A generalizability analysis of score consistency for the Balanced Inventory of Desirable Responding. *Psychological assessment*, *25*(1), 94-104.

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Thousand Oaks: Sage.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished Doctoral dissertation, University of California, Los Angeles.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*(1), 115-148.

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12*(2), 250-264.

Heimildir

Albright, J. J. og Park, H. M. (2009). Confirmatory factor analysis using Amos, LISREL, Mplus, and SAS/STAT CALIS. *Technical working paper.* Bloomington, IL: Indiana University.

Baker, F.B. (2001). *The basics of item response theory.* USA: ERIC Clearinghouse on Assessment and Evaluation.

Barrick, M. R. og Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44,* 1-26.

Barrick, M. R. og Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261-272.

Beatty, P. C. og Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287-311.

Behling, O. og Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. London: SAGE Publications.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246.

Beretvas, S. N., Meyers, J. L. og Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *62*(4), 570-589.

Bollen, K. A. (1990). Overall fit in covariance structure models: two types of sample size effects. *Psychological Bulletin*, *107*(2), 256-259.

Booth-Kewley, S., Edwards, J. E. og Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77*(4), 562-566.

Browne, M. W. og Cudeck, R. (1993). Alternative ways of assessing model fit. Í K. A. Bollen og J. S. Long (ritstjórar), *Testing structural equation models* (pp. 136-136). Newbury Park, CA: Sage Publications, Inc.

Cai, L., Du Toit, S. H. C. og Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [hugbúnaður]. *Chicago, IL: Scientific Software International*.

Cervellione, K. L., Lee, Y. S. og Bonanno, G. A. (2009). Rasch modeling of the self-deception scale of the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *69*(3), 438-458.

Clark, L. A. og Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309-319.

Costa, P. T. og MacCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI) professional Manual*. Odessa, FL: Psychological Assessment Resources.

Crowne, D. P. og Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349-354.

DeVellis, R. F. (2003). *Scale development: Theory and applications.* Thousand Oaks: Sage.

Edelen, M. O. og Reeve B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16* (1), 5-18.

Einar Guðmundsson (2005-2006). Þýðing og staðfærsla sálfræðilegra prófa. *Sálfræðiritið – Tímarit Sálfræðingafélags Íslands, 10 – 11,* 23-40.

Flora, D. B. og Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*(4), 304-312.

Helmes, E. og Holden, R. R. (2003). The construct of social desirability: one or two dimensions? *Personality and Individual Differences*, *34*, 1015–1023.

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science, 39*(3), 184-201.

Holden, R. R., Starzyk, K. B., McLeod, L. D. og Edwards, M. J. (2000). Comparisons among the Holden Psychological Screening Inventory (HPSI), the Brief Symptom Inventory (BSI), and the Balanced Inventory of Desirable Responding (BIDR). *Assessment*, *7*(2), 163-175.

Hu, L. T. og Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

Hu, L. T. og Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Hu, L. og Bentler, P. M. (1995). Evaluating model fit. Í R. H. Hoyle (ritstjóri), *Structural equation modeling: Issues, concepts, and applications* (pp. 76-99). Newbury Park, CA: Sage.

International Test Commision (2000). *ITC Guidelines on Adapting Tests.* Sótt 11. desember 2013 af http://www.intestcom.org/upload/sitefiles/40.pdf

Jones, E. E. og Sigall, H. (1971). The bogus pipeline: a new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*(5), 349-364.

Jöreskog, K. G. og Sorbom, D. (2012). LISREL 9.1 [hugbúnaður]. *Lincolnwood, IL: Scientific Software International*.

Jöreskog, K.G. (1993). Testing Structural Equation Models. Í Bollen, K.A. og Long, S.J. (ritstjórar), *Structural equation models* (bls. 295-316). California: SAGE Publications, Inc.

Kam, C. (2013). Probing item social desirability by correlating personality items with Balanced Inventory of Desirable Responding (BIDR): A validity examination. *Personality and Individual Differences,54,* 513-518.

Lalwani, A. K., Shavitt, S. og Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding?. *Journal of Personality and Social Psychology*, *90*(1), 165-178.

Lalwani, A. K., Shrum, L. J. og Chiu, C. Y. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology*, *96*(4), 870-882.

Lanyon, R. I. og Carle, A. C. (2007). Internal and external validity of scores on the Balanced Inventory of Desirable Responding and the Paulhus Deception Scales. *Educational and Psychological Measurement*, *67*(5), 859-876.

Leite, W. L. og Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne social desirability scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *65*(1), 140-154.

Li, A. og Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment*, *14*(2), 131-141.

Li, A. og Bagger, J. (2007). The Balanced Inventory of Desirable Responding (BIDR): A Reliability Generalization Study. *Educational and Psychological Measurement*, *67*(3), 525-544.

Li, A. og Reb, J. (2009). A cross-nations, cross-cultures, and cross-conditions analysis on the equivalence of the Balanced Inventory of Desirable Responding. *Journal of Cross-Cultural Psychology*, *40*(2), 214-233.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*(5), 493.

Mallinson, S. (2002). Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. *Social Science & Medicine*, *54*(1), 11-21.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, *42(*5*)*, 851–858.

Mount, M. K., Barrick, M. R. og Stewart, G. L. (1998). Personality predictors of performance in jobs involving interaction with others. *Human Performance*, *11*(3), 145-166.

Nunnally, J. C. og Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Ones, D. S. og Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*(2-3), 245-269.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598-609.

Paulhus, D. L. (1991). Measurement and control of response bias. Í J. P. Robinson, P. R. Shaver og L. S. Wrightsman (ritstjórar), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L. (1994). Balanced inventory of desirable responding: Reference manual for BIDR version 6. Óbirt handrit, *University of British Columbia, Vancouver, Canada*.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. Í H. I. Braun, D. N. Jackson og D. E. Wiley (ritstjórar), *The role of constructs in*

psychological and educational measurement (bls. 49-69). Mahwah NJ: Lawrence Erlbaum Associates, Inc.

Paulhus, D. L., Bruce, M. N. og Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, *21*(2), 100-108.

Pauls, C. A. og Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*(6), 1137-1151.

Pauls, C. A. og Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences*, *35*(2), 263-275.

Presser, S. og Blair, J. (1994). Survey pretesting: Do different methods produce different results. *Sociological Methodology*, *24*(1), 73-104.

Raykov, T. og Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling*, *13*(1), 130-141.

Reid-Seiser, H. L. og Fritzsche, B. A. (2001). The usefulness of the NEO PI-R Positive Presentation Management Scale for detecting response distortion in employment contexts. *Personality and Individual Differences*, *31*(4), 639-650.

Reise, S. P. og Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*(2), 93-103.

Ritchie, J., Spencer, L. og O'Connor, W. (2003). Carrying out qualitative analysis. *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, 219-262.

Rogelberg, S. G., Conway, J. M., Sederburg, M. E., Spitzmüller, C., Aziz, S. og Knight, W. E. (2003). Profiling active and passive nonrespondents to an organizational survey. *Journal of Applied Psychology*, *88*(6), 1104-1114.

Sackeim, H. A. og Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47*(1), 213-215.

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*,*76*(3), 323-346.

Sax, L. J., Gilmartin, S. K. og Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, *44*(4), 409-432.

Sharma, S., Mukherjee, S., Kumar, A. og Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*(7), 935-943.

Shevlin, M. og Miles, J. N. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, *25*(1), 85-90.

Smith, M. A. og Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 496-505.

Stanton, J. M., Sinar, E. F., Balzer, W. K. og Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*(1), 167-194.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173-180.

Stöber, J., Dette, D. E. og Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment*, *78*(2), 370-389.

Háskóli Íslands. (2013). *HÍ í tölum*. Sótt 14. desember, 2013, af https://docs.google.com/viewer?url=http%3A%2F%2Fwww.hi.is %2Fsites%2Fdefault%2Ffiles%2Fskradir_feb_2013_heild.xlsx

Vispoel, W. P. og Tao, S. (2013). A generalizability analysis of score consistency for the Balanced Inventory of Desirable Responding. *Psychological Assessment*, *25*(1), 94-104.

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Thousand Oaks: Sage.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Óbirt Doktorsritgerð, University of California, Los Angeles.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*(1), 115-148.

Zerbe, W. J. og Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12*(2), 250-264.