

Tæknivædd prófatriði:

Tæknilegir eiginleikar,
fjölbreytni og framboð

Bára Elísabet Dagsdóttir
Sigurgrímur Skúlason
Nói Kristinsson
Ágúst 2021

**Tæknivædd prófatriði: Tæknilegir eiginleikar,
fjölbreytni og framboð**

ISBN 978-9979-0-2745-4

©2021 Bára Elísabet Dagsdóttir, Sigurgrímur Skúlason og Núi Kristinsson

Málfarslestur: Berglind Steinsdóttir

Öll réttindi áskilin

1. útgáfa 2021 rafræn

Menntamálastofnun

Kópavogi

Hönnun og umbrot: Menntamálastofnun

Þetta verkefni var unnið með styrk frá Nýsköpunarsjóði námsmanna sumarið 2021.

Höfundar þakka RANNÍS veittan stuðning.

Efnisyfirlit

Inngangur	4
Ólíkar tegundir TVPA	7
Atriði þar sem próftakar velja hlut (e. <i>choice items</i>)	10
Atriði þar sem próftakar tengja eða raða hlutum (e. <i>match & order</i>)	11
Atriði þar sem próftakar bæta við hlut (e. <i>pre-constructed completion</i>).	11
Atriði þar sem próftakar vinna inni í hlut (e. <i>custom interactive</i>).	12
Atriði þar sem próftakar gera eða útbúa hlut (e. <i>free construction</i>).	13
Atriði þar sem próftakar vinna með hlut (e. <i>upload</i>).	14
Prófakerfi	14
Learnosity	14
Inspira	15
TAO	15
Assessment Systems (Ava og FastTest)	15
Ava	15
FastTest	15
Wiseflow/Uniwise	16
SoNET	16
Rannsóknir	18
Samantekt	19
Heimildir	21

Inngangur

Prófatriði eru grunneiningar allra prófa og veita upplýsingar um kunnáttu, leikni eða færni próftaka á tilgreindu sviði (Haladyna og Rodriguez, 2013; Lane og Iwatani, 2016; Millman og Greene, 1989; Osterlind, 1997; Schmeiser og Welch, 2006; Tinkelman, 1971). Hvert og eitt prófatriði gefur próftaka kost á að sýna að hvað marki hann hefur náð valdi á tiltekinni kunnáttu, leikni eða færni. Þau eru byggð þannig upp að hvert og eitt þeirra gefur próftökum, hvort sem eru með meiri eða minni færni, færi á að sýna hvað þeir geta. Því er nauðsyn að í prófi séu nokkur prófatriði sem reyna á sömu eða svipaða eiginleika en hvert og eitt þeirra gefi próftökum með mismikla færni kost á að sýna hvaða þeir geta. Þar sem prófatriðið reynir einungis á þrönga hlið þeirrar kunnáttu, leikni og færni sem prófið metur þurfa að vera mörg prófatriði í hverju prófi til að það nái að gefa víðtækar upplýsingar (Haladyna og Rodriguez, 2013; Lane og Iwatani, 2016; Schmeiser og Welch, 2006). Þá getur líka verið æskilegt að próf innihaldi ólíkar gerðir prófatriða sem gefa próftökum kost á að nálgast viðfangsefni með ólíkum hætti, bæði ólíkar tegundir prófatriða og ólík afbrigði af sömu tegund prófatriða.

Fjöldamörg orð eru notuð til að vísa til hegðunar, hæfni, kunnáttu, reynslu, skoðana, tilfinninga eða viðhorfa sem matstæki leggja mat á. Í þessari skýrslu verður orðið færni notað almennt um þá eiginleika sem meta skal. Orðið hæfni samsvarar enska orðinu „ability“ sem oftast er notað í þessum tilgangi í próffræðilegri umfjöllun um mælitæki.

Almennt séð eru prófatriði byggð upp af tveim hlutum, spurnarlið og svarlið (Haladyna og Rodriguez, 2013; Lane og Stone, 2006; Roid og Haladyna, 1982). Spurnarliður er sá hluti prófatriðis þar sem viðfangsefnið er sett fram. Svarliður er sá hluti þess sem gefur próftaka kost á að bregðast við eða leysa viðfangsefnið sem sett er fram í spurnarliðnum (Haladyna og Rodriguez, 2013; Osterlind, 1997; Schmeiser og Welch, 2006). Hefðbundnum prófatriðum má svo skipta í tvær megintegundir atriða (e. *item formats*); atriði þar sem próftaki velur svar (e. *selected response*) og atriði þar sem próftaki setur sjálfur upp svarið (e. *constructed response* eða *performance assessment*). Algengasta gerð atriða þar sem svar er valið úr mengi nokkurra möguleika er fjölvalsspurningar, sem er sú gerð prófatriða sem er jafnframt mest rannsökuð. Aðrar algengar tegundir atriða af þessu tagi eru tengiatriði (e. *mapping*), þar sem próftaki tengir stök í einu mengi við stök í öðru mengi, eða atriði þar sem próftakar raða eða flokka stök í mengi. Atriði sem krefjast þess að próftaki útbúi eða setji saman svar, samsett prófatriði, eru fjölbreyttari, t.a.m. opnar spurningar, ritgerðarspurningar og atriði þar sem próftakar fylla í eyður (Bennett, 1993; Fitzpatrick og Morrison, 1971; Haladyna og Rodriguez, 2013).

Í kjölfar tækniframfara síðustu áratuga eru sífellt fleiri próf lögð fyrir rafrænt eða í stafrænu umhverfi. Stafrænar fyrirlagnir bjóða upp á nýjar gerðir prófatriða, eða nýjar útfærslur eldri gerða af prófatriðum, sem hafa verið kölluð tæknivædd prófatriði eða TVPA (e. *technology-enhanced items*; TVPA) (Bennett, 2003; Schechinger, 2012; Sireci, og Zenisky, 2006; Swiggett, 2014; Wan og Henly, 2012; Wise., Soland og Dupray, 2021; Zenisky og Stephen, 2002). TVPA nýta möguleika stafræns umhverfis til gagnvirkni (e. *interactions*) milli próftaka og prófakervis eða prófatriðis (Bennett, 2002; Haladyna og Rodriguez, 2013). Gagnvirkni próftaka og prófatriðis er almennt álitin flóknari í TVPA en í hefðbundnum prófatriðum þar sem TVPA gefa próf-

taka kost á að búa til eða hagræða efni til að svara atriðum (Russell, 2016). Til eru fjölmargar gerðir TVPA þar sem stöðug tæknileg þróun býður upp á nýja möguleika.

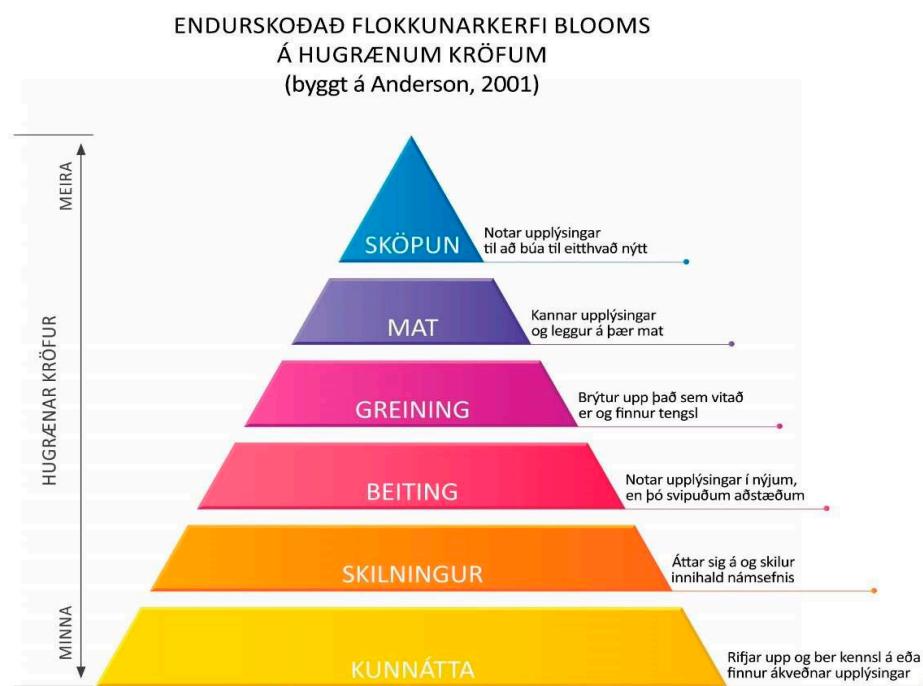
Það getur reynst flókið að ákvarða hvers konar atriðategundir henta best hverju sinni. Eftir að rafrænar fyrirlagnir komu til sögunnar er úr mun fleiri tegundum prófatriða að velja. Í umfjöllun sinni um prófatriði telja Haladyna og Rodriguez (2013) upp nokkra þætti sem gott er að hafa í huga við val á tegundum prófatriða. Í fyrsta lagi þarf að huga að því hvað færni sem svarliður atriðategundar krefst af próftaka svipar mikið til viðmiðunarhegðunar (e. *criterion behaviour*), þ.e. hvernig einstaklingar beita hæfninni í raunverulegum aðstæðum. Í öðru lagi þarf að meta hve flóknar hugrænar kröfur atriðategundin gerir á próftaka (e. *cognitive complexity*). Prófatriði geta verið breytileg að þessu leyti á tvo vegu, annars vegar með tilliti til leiðbeininga til próftaka og skilyrða eða forsenda fyrir svörum hins vegar. Til að mynda eru fyrir-mæli fyrir svörum fjölvalsspurninga almennt stutt og tiltölulega einföld, en leiðbeiningar og svörum opnari spurninga eins og ritgerðarspurninga geta verið breytilegri. Mismunandi tegundir prófatriða hafa misjafna uppbyggingu og því er mikilvægt að skilja hvaða færni og hve flókna hugræna færni er hægt að mæla með hverri tegund. Fræðimenn hafa bent á nokkrar mismunandi víddir sem prófatriði eru breytileg á (Haladyna og Downing, 2004; Haladyna og Rodriguez, 2013; Messick, 1989; 1994; Rodriguez, 2002):

1. Hlutlægni eða huglægni matskvarða sem skorað er út frá (e. *objective vs. subjective scoring*)
2. Eðli svarliða, valmöguleikar eða aðgerðir (e. *selection vs. production*)
3. Staðlaður eða opinn svarliður (e. *fixed-response vs. free-response*)
4. Metinn er afurð (afrakstur) eða frammistaða (ferli) (e. *product vs. performance*)

Með hlutlægri eða huglægri skorun er átt við hve skýr og afdráttarlaus stigagjöf fyrir svör við atriðinu er. Fjölvalsspurningar hafa yfirleitt aðeins eitt rétt svar, auðvelt er að meta hvort próftaki svarar slíkum atriðum rétt. Skorun annarra atriðategunda getur svo verið huglægari, þar sem aðrir einstaklingar þurfa að lesa yfir svör próftaka og meta gæði þeirra, oft með hjálp einhvers konar matskvarða, en þetta fyrirkomulag býður upp á meiri villu í stigagjöf. Val eða samsetning felur svo í sér hvað próftaki þarf að gera til að svara atriðinu. Í atriðum eins og fjölvalsspurningum þarf próftaki að meta svarlið og svarmöguleika og velja síðan þann svarlið sem best samsvarar niðurstöðunni, en í atriðum með samsett svör þarf próftaki sjálfur að búa til og setja saman svar sitt. Stöðlun eða sveigjanleiki (e. *fixed vs. free-response*) er vídd sem á við um atriði þar sem próftaki setur sjálfur saman svar, en þau geta verið breytileg eftir því hve mikið frelsi próftaki hefur til að tjá sig. Ef atriði hafa staðlaða svarstiku (e. *fixed-response*) hefur próftaki minna frelsi og fyrirmæli kalla á ákveðið svar. Frjálsari spurningar gefa próftökum meira svigrúm í svörum. Í sumum samsettum atriðategundum er svo áherslan á afurð en í öðrum er einblínt meira á frammistöðu eða tæknina sem próftaki beitir til að svara spurningunni. Dæmi um mat þar sem áhersla er á afurð er ritgerð eða smíðisgripur. Dæmi um mat þar sem áhersla er á frammistöðu væri mat á dansnema eða framsögn og framkomu í ræðustól í ræðukeppni eða þegar kjósendur gera upp hug sinn milli frambjóðenda á kosningafundi. Dæmi um aðstæður þar sem bæði afurð og frammistaða eru metin eru mat á veitingahúsi, þar sem bæði er metin afurð, maturinn, og frammistaða, framkoma þjóna.

Eitt meginmarkmið TVPA er að gefa próftökum færi á að sýna vald sitt á flókinni og breytilegri kunnáttu, leikni eða færni, sem erfitt er að meta með hefðbundnum atriðum (Beller, 2013). Dæmi um stigvaxandi flókna færni má sjá á myndrænni framsetningu á endurbættu flokkunarkerfi Blooms, sjá mynd 1. Samkvæmt flokkunarkerfinu, sem hefur verið notað til að flokka og þróa hæfniviðmið fyrir nám, skiptist hugræn færni í sex þekkingarsvið, þar sem færnin verður flóknari í efri flokkum. Kerfið má svo nota til að flokka prófatriði eftir því hvaða þekkingarsvið próftakar nota við svörun (Anderson o.fl. 2001; Krathwohl, 2002). Vonir standa til þess að TVPA geti prófað færni á efri þekkingarsviðum og þannig veitt upplýsingar um flóknari hugræna færni en hefðbundin prófatriði (Beller, 2013; Bennett, 2002; Murray, e.d.; Russell og Moncaleano, 2019; Wise., Soland og Dupray, 2021).

Mynd 1. Endurskoðað flokkunarkerfi Blooms fyrir hugrænar kröfur viðfangsefna í námi (byggt á Anderson, 2001).



Annar hugsanlegur kostur við TVPA, sem er drifkraftur að baki þróun flestra tegunda slíkra prófatriða, er möguleikinn á að úrlausn eða svarliður slíkra prófatriða gæti verið nær raunverulegri viðmiðunarhegðun en kostur er í úrlausn eða svarliðum hefðbundinna prófatriða á pappír. Hefðbundin prófatriði sem útfærð eru á pappír, svo sem fjölvalsspurningar, hafa verið gagnrýnd á þeim grundvelli að úrlausn þeirra eigi sér ekki hliðstæðu í raunverulegri hegðun. Þar af leiðandi meti þau ekki færni sem fólk nýtir sér í daglegu lífi. TVPA bjóða upp á fjölbreyttari úrlausnarmöguleika, sem hægt er að hanna til að líkjast lausnum raunverulegra verkefna, og því standa vonir til þess að TVPA geti hjálpað til við að leysa þennan vanda (Sireci og Zenisky, 2006; Murray, e.d.). Loks hafa ýmsir haldið því fram að TVPA styrki réttmæti niðurstaðna á prófum vegna þess að inntak þeirra eða efnislegt innihald endurspegli betur þá hugsmíð sem prófinu er ætlað að meta. Orðið innihaldsréttmæti er oft notað í þessu samhengi en fræðilegt sjónarhorn próffræðinnar á réttmæti hefur þróast mjög á undanförunum áratugum. Nútíma-

kenningar um réttmæti líta svo á að það, hversu vel prófatriði endurspegla hugsmíð að baki prófinu, sé hluti af þeim gögnum sem huga þarf að þegar réttmæti túlkunar á þeim eru metin (Cizek, 2020; Kane, 2003, 2013; Sireci og Zenisky, 2006). Að sumra mati er unnt að nálgast færni sem prófinu er ætlað að meta betur með TVPA og á nýjan og fjölbreytilegri hátt sem erfitt hefur reynst að meta með hefðbundnum prófatriðum (t.d. Parshall o.fl., 2010; Sireci og Zenisky, 2006). Hins vegar skortir rannsóknir á eiginleikum TVPA sem styðja þessar fullyrðingar (Wan og Henly, 2012).

Fræðimenn hafa einnig bent á mögulega galla eða veikleika við TVPA. TVPA gætu verið viðkvæm fyrir áhrifum breytileika sem er ótengdur hugsmíðinni (e. *construct irrelevant variance*) sem ætlunin er að mæla (Cizek, 2020; Kane, 2006, 2013; Messick, 1975, 1989, 1996). Það liggur beint við að ætla að tölvufærni próftaka hafi áhrif á getu þeirra til að svara atriðunum en fleira getur haft áhrif (Bennett, Morley og Quardt, 1998; Hambleton, 2002; Zenisky og Stephen, 2002; Way, Steffen, Anderson, 2002). Tölvufærni gæti verið óviljandi metin með TVPA sem ætti, ef allt væri eins og best yrði á kosið, einungis að mæla kunnáttu, leikni eða færni á tilgreindu efnissviði. Allar aðferðir við prófun takmarkast af einkennum aðferðarinnar, t.d. próf á pappír með fjölvals- eða opnum spurningum eða munnleg próf fyrir framan mismarga prófdómara. Í margreyndum aðferðum eru þessi áhrif þekkt, en þau eru líka mismikil, og ein ástæða þess hve mikið fjölvalsspurningar eru notaðar er sú að þær reyndust hafa minni aðferðaráhrif en t.d. ritgerðarspurningar eða munnleg próf. Breytileiki ótengdur hugsmíðinni felst í því að prófatriði eða aðferð við prófun krefjist það mikillar færni af öðrum toga en hugsmíðin felur í sér að hún hafi trúlandi áhrif á ályktanir sem felast í túlkun niðurstaðna (Messick, 1989; Parshall, 2002). Þetta getur birst í tengslum við TVPA með þeim hætti að áhugi nemenda beinist meira að því að fikta í tæknimöguleikum og að prófa allt sem hægt er að fikta í fremur en að leysa verkefnið markvisst með þeim aðgerðum sem til þarf. Þannig þarf aukinn áhugi ekki endilega að leiða til þess að geta hvers nemanda sé metin með betri hætti. Einnig getur það hvernig tæki nemendur nota haft áhrif á það hve auðvelt er að leysa sum tæknivædd prófatriði. Þannig getur það í reynd verið gjörólíkt viðfangsefni að endurraða hlutum á skjá, sem gerist t.a.m. í hreyfi-prófatriðum (e. *drag-and-drop*), með mús en með bendingu á skjá. Einnig getur það verið mjög ólíkt verkefni að setja inn staðsetningu í staðsetningarprófatriðum (e. *hot-spot items*) eftir því hvaða tækni er notuð.

Ólíkar tegundir TVPA

Til eru fjölmargar gerðir TVPA en flokkun þeirra hefur reynst erfið þar sem sífellt fleiri tegundir bætast við (Hambleton, 2002; Hamer og Jongkamp, 2021; Parshall, 2002). Tvær tilraunir til að flokka TVPA eftir eiginleikum þeirra hafa verið gerðar. Scalise (Scalise og Gifford, 2006) próaði kerfi sem byggir á tveim víddum, annars vegar hve sveigjanlegur svarliður prófatriðis er og hins vegar hve flókin tæknileg útfærsla prófatriðis er. Á síðustu árum hafa Hamer og Jongkamp (2021) reynt að styrkja fræðilegar stoðir flokkunar TVPA með endurskoðun á kerfi Scalise þar sem notaðar eru víddirnar sveigjanleiki svarliðar og hugrænar kröfur sem prófatriði gerir til próftaka. Áður höfðu Parshall og félagar (Parshall, Davey, & Pashley, 2000; Parshall, Harnes,

Davey, og Pashley, 2010) sett fram flokkunarkerfi þar sem unnið var með sjö einkenni TVPA. Kerfi þeirra lýtur að mun tæknilegri víddum eða einkennum prófatriða en kerfi Scalise eða Hamer og Jongkamp. Kerfi Parshall og félaga getur hentað sérfræðingum sem vinna að þróun TVPA en er óþarflega flókið til að draga fram heildarmynd og kemur í sumum tilvikum inn á einkenni prófatriða sem litlu skipta í heildarmyndinni. Því byggir umfjöllun hér fremur á heildarmynd sem fram kemur hjá Scalise og hjá Hamer og Jongkamp en á tæknilegum atriðum sem Parshall og félagar vinna með.

Í flokkun Scalise á TVPA eru notaðar tvær víddir. Annars vegar sveigjanleiki prófatriða og hins vegar hve flókin þau eru. Það er síðari víddin sem Hamer og Jongkamp (2021) nota ekki. Scalise og Gifford telja að hefðbundnar tegundir prófatriða væru annað hvort algjörlega skorðaðar (e. *fully constrained*), eins og fjölvalsspurningar, eða væru algjörlega samsettar (e. *fully constructed*), eins og ritgerðarspurningar. Nýjar atriðagerðir TVPA liggi á milli þessara tveggja póla. Eins og aðrir fræðimenn bundu Scalise og Gifford vonir við að TVPA gætu þannig haft þann kost að meta flókna hugræna færni, líkt og samsett atriði gera, án þeirra takmarkana sem geta fylgt algjörlega skorðuðum atriðum eins og fjölvalsspurningum (Scalise og Gifford, 2006).

Seinni víddin felur svo í sér hve flókinni færni próftaki þarf að beita til að leysa prófatriðin. Þessi vídd virðist ekki eins mikilvæg í flokkunarkerfi Scalise og Gifford og í raun er lítið fjallað um hana í grein þeirra um kerfið (Scalise og Gifford, 2006). Hamer og Jongkamp bentu á að þessi vídd hjá Scalise og Gifford sneri meira að því hve flókin hönnun atriðanna er, frekar en hve flóknar hugrænar kröfur úrlausn atriðanna setur (Hamer og Jongkamp, 2021). Á mynd 2 má sjá myndræna framsetningu á flokkunarkerfi Scalise og Gifford. Athyglisvert er að hér er engin skilgreining á því hvað felst í meira og minna lóðréttu víddinni.

Tafla 1. Tegundir tæknivæddra prófatriða samkvæmt flokkunarkerfum Scalise (Scalise og Gifford, 2006) og Hamer og Jongkamp (2021).

Flokkunarkerfi			
Hamer og Jongkamp, 2019		Scalise, 2006, 2009	
Íslenska	Enska	Íslenska	Enska
Velja hlut	Choice	Fjölval	Multiple-choice
		Val og auðkenning	Selection/identification
Tengja eða raða hlutum	Match & order	Röð og skipulag	Reorder/rearrangement
Bæta við hlut	Pre-construct completion	Viðbætur og breyting	Substitution/correct
Vinna inni í hlut	Custom interactive	Innsetning	Completion
Gerð hluta	Free construction	Gerð hluta	Construction
Vinna með hlut	Upload	Kynning	Presentation

Mynd 2. Flokkunarkerfi Scalise fyrir TVPA (Scalise og Gifford (2006) og Scalise (2012)).

Most Constrained							Least Constrained
Less Complex	Fully Selected	Intermediate Constraint Item Types					Fully Constructed
	1. Multiple Choice	2. Selection/ Identification	3. Reordering/ Rearrangement	4. Substitution/ Correction	5. Completion	6. Construction	7. Presentation/ Portfolio
	1A. True/False (Haladyna, 1994c, p.54)	2A. Multiple True/False (Haladyna, 1994c, p.58)	3A. Matching (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. Interlinear (Haladyna, 1994c, p.65)	5A. Single Numerical Constructed (Parshall et al, 2002, p. 87)	6A. Open-Ended Multiple Choice (Haladyna, 1994c, p.49)	7A. Project (Bennett, 1993, p.4)
	1B. Alternate Choice (Haladyna, 1994c, p.53)	2B. Yes/No with Explanation (McDonald, 2002, p.110)	3B. Categorizing (Bennett, 1993, p.44)	4B. Sore-Finger (Haladyna, 1994c, p.67)	5B. Short-Answer & Sentence Completion (Osterlind, 1998, p.237)	6B. Figural Constructed Response (Parshall et al, 2002, p.87)	7B. Demonstration, Experiment, Performance (Bennett, 1993, p.45)
	1C. Conventional or Standard Multiple Choice (Haladyna, 1994c, p.47)	2C. Multiple Answer (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. Ranking & Sequencing (Parshall et al, 2002, p.2)	4C. Limited Figural Drawing (Bennett, 1993, p.44)	5C. Cloze-Procedure (Osterlind, 1998, p.242)	6C. Concept Map (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. Discussion, Interview (Bennett, 1993, p.45)
More Complex	1D. Multiple Choice with New Media Distractors (Parshall et al, 2002, p.87)	2D. Complex Multiple Choice (Haladyna, 1994c, p.57)	3D. Assembling Proof (Bennett, 1993, p.44)	4D. Bug/Fault Correction (Bennett, 1993, p.44)	5D. Matrix Completion (Embretson, S, 2002, p. 225)	6D. Essay (Page et al, 1995, 561-565) & Automated Editing (Breland et al, 2001, pp.1-64)	7D. Diagnosis, Teaching (Bennett, 1993, p.4)

Hamer og Jongkamp (2021) hafa unnið að því undanfarin ár að búa til flokkunarkerfi fyrir TVPA, sem er innblásið af eldra og að miklu leyti úreltu kerfi Scalise (2012; Scalise o.fl., 2021; Scalise og Gifford, 2006). Í flokkunarkerfi Hamer og Jongkamp (2021) eru prófatriði flokkuð eftir tveimur víddum. Annars vegar hve sveigjanlegur svarliður atriðisins er og hins vegar hve miklar eða flóknar hugrænar kröfur prófatriði gerir til próftaka. Fyrri víddin liggur með vaxandi sveigjanleika frá því að próftaki velji á milli eða flokki hluti í stafræna umhverfinu yfir í að próftaki vinni með eða noti flókna stafræna hluti. Einföldustu TVPA eiga sér samsvörun í prófatriðum í pappírsumhverfi og gefa próftaka kost á að velja milli hluta í stafrænu umhverfi (e. *selection items*). Þetta eru stafrænar útfærslur af fjölvalsspurningum og tvíkosta valprófatriðum. Ívið meiri sveigjanleiki er í TVPA þar sem próftaki skilgreinir tengsl milli hluta eða flokkar þá (e. *match & order*). Meiri sveigjanleiki felst í prófatriðum sem krefjast þess að próftaki bæti inn í það sem sett er fram í prófatriðinu (e. *pre-constructed completion*), enn meiri sveigjanleiki felst í prófatriðum sem krefjast þess að próftaki vinni með hlut í stafræna

umhverfinu (e. *custom interactive*) og skapi sjálfur hlut í stafræna umhverfinu (t.d. texti, hljóð-upptaka af munnlegu svari, teikning; e. *free construction*). Sveigjanlegasta tegund TVPA krefst þess að próftakar vinni á einhvern hátt með framsetningu stafrænna hluta eða breyti þeim eða þrói þá áfram (t.d. með hljóðskrá, myndskrá, forriti eða textaskjali; e. *upload*).

Seinni víddin er e.t.v. óljósari og ekki eins skýrt útfærð í flokkunarkerfi Hamer og Jongkamp. Þessi vídd reynir að skilgreina stigvaxandi hugrænar kröfur sem prófatriði krefjast af próftökum við úrlausn viðfangsefnis prófatriðisins. Þessi vídd byggir á flokkunarkerfi Blooms (Anderson o.fl. 2001). Í kerfi Hamer og Jongkamp er aukning í hugrænum kröfum, þ.e. aukning á seinni víddinni, ekki jöfn í öllum tilvikum frá einum sveigjanleikaflokki eða stigi til þess næsta, þ.e. flokkar fyrri víddarinnar. Til dæmis er tiltölulega lítil breyting á sveigjanleika milli prófatriða sem gefa próftaka kost á að velja hlut annars vegar og prófatriða þar sem próftaki á tengja hins vegar, eða á milli atriða þar sem krafist er tengingar annars vegar og prófatriða þar sem próftaki á að raða hlutum hins vegar. Mun meiri aukning er í sveigjanleika sem birtist í mismun á því að vinna inn í hlut yfir í að gera hlut. Samspil þessara tveggja vídda er flókin. Unnt er að þróa prófatriði sem hafa lítinn sveigjanleika þannig að þau geri ýmist litlar hugrænar kröfur eða miklar hugrænar kröfur. Á sama hátt geta prófatriði með meiri sveigjanleika bæði gert miklar og litlar hugrænar kröfur til próftaka. Í sumum flokkum eru þau prófatriði sem gera mestu hugrænu kröfurnar töluvert einfaldari að gerð en þau sem gera minni kröfur. Til að mynda er samsett val (e. *composite choice*), sú tegund prófatriða sem gerir mestar hugrænar kröfurnar í minnst sveigjanlega flokknum (e. *choice items*), mun einfaldara en kennsla (e. *teaching, coaching*), sem er atriðið með mestu hugrænu kröfurnar í sveigjanlegasta flokknum (e. *upload*) (Hamer og Jongkamp, 2021).

Hér að neðan er nánari umfjöllun um hvern flokk og útskýring á atriðum sem falla í þá. Umfjöllunin um flokkana er í röð frá þeim með minnst sveigjanlega svarliði (þ.e. *choice items*) til þess sem er mest sveigjanlegur (þ.e. *upload*). Umfjöllun um atriði innan hvers flokks er svo í röð eftir því hve flóknar hugrænar kröfur atriðin gera til próftaka. Þar að auki fylgja með myndir af TVPA úr öllum flokkum sem skilgreindir eru í kerfinu, sem eru fengnar frá prófakerfi *Learnosity*.

Atriði þar sem próftakar velja hlut (e. *choice items*)

Þessi flokkur prófatriða á sér samsvörun í stöðluðum prófatriðum sem þróuð hafa verið fyrir próf á pappír. Fjölmargar tegundir fjölvalsprófatriða og tvíkosta rétt/rangt prófatriða hafa verið útfærðar. Þau eru til sem stök prófatriði, prófatriði með mynd eða myndskaiði í spurnarlið, sem fylki eða hópur prófatriða sem sett eru upp í sömu skjámynd og í fleiri útfærslum. Einfaldasta tegundin í þessum flokki eru atriði þar sem próftaki velur annað svarið af tveimur, næst koma atriði þar sem valið er úr þremur eða fleiri svörum, atriði þar sem fyllt er í eyður með því að velja t.d. rétt orð, atriði þar sem fleiri en einn svarmöguleiki eru réttir og loks atriði þar sem próftaki velur svar og þarf svo að útskýra svarið. Þessi flokkur samsvarar tveim flokkum atriða í kerfi Scalise, útfylling (e. *fill in the blanks/cloze*) og flokkun, pörun og röðun (e. *classify, match & order*), líkt og kemur fram í töflu 1.

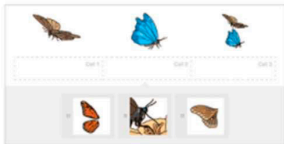
Mynd 3. Dæmi um atriði frá Learnosity þar sem próftakar velja hlut (Murray (e.d.)).

Multiple choice – standard	Multiple choice – multiple response	True or false
Which color has the smallest wavelength? ☐ Red ☒ Violet ☐ Green	Which color has the smallest wavelength? ☒ Red ☐ Violet ☒ Green	The sky is blue due to gases. ☒ True ☐ False

Atriði þar sem próftakar tengja eða raða hlutum (e. *match & order*)

Til að svara þessum tegundum TVPA þarf próftaki að tengja saman eða raða hlutum, hugtökum eða annars konar hugsmíðum. Einfaldasta tegundin úr þessum flokki eru atriði þar sem próftaki tengir saman pör, þ.e. tvo hluti (e. *one-on-one match*), næst kemur flokkun þar sem próftakar flokka hluti í viðeigandi flokka, þar á eftir er röðun þar sem próftakar raða hlutum eftir einni vídd, svo uppbygging (e. *structuring*) þar sem próftakar setja saman hluti eftir ákveðnu mynstri sem er flóknara en einföld röðun. Loks er flóknasta tegund prófatriða í þessum flokki röðun og endurröðun (e. *arrange and rearrange*) þar sem próftakar setja aftur saman hluti eftir flóknara mynstri en einfaldri röðun, en þurfa auk þess að gera grein fyrir tengslunum milli þeirra.

Mynd 4. Dæmi um atriði frá Learnosity þar sem próftakar tengja eða raða hlutum (Murray (e.d.)).

Classification	Match list	Order list
		

Atriði þar sem próftakar bæta við hlut (e. *pre-constructed completion*).

Þessi tegund prófatriða felur í sér að próftakar bæti við eða klári það sem sett er fram í atriðinu. Einfaldasta tegundin í þessum flokki gengur út á að próftaki gefi svar sem samanstendur af tölu, stöfum eða fáeinum orðum, eða einhvers konar einfaldri samsetningu af þessu þrennu, þ.e. stutt, samsett svar. Næst koma atriði þar sem próftaki þarf að teikna einföld form, svo sem línur, þríhyrning eða hring. Því næst eru atriði þar sem próftaki færir til, hagræðir eða snýr einföldum formum. Svo koma atriði þar sem próftakar velja, færa eða snúa formum sem gefin eru upp í valmynd til þess að búa til mynstur, graf eða skýringarmynd. Loks krefjast flóknustu atriðin í þessum flokki þess að próftaki velji, færi, snúi eða hagræði formum úr valmynd, teikni ný form eða tengingar og búi þannig til mynstur, flæðirit eða skýringarmynd.

Mynd 5. Dæmi um atriði frá Learnosity þar sem próftakar bæta við hlut (Murray (e.d.)).

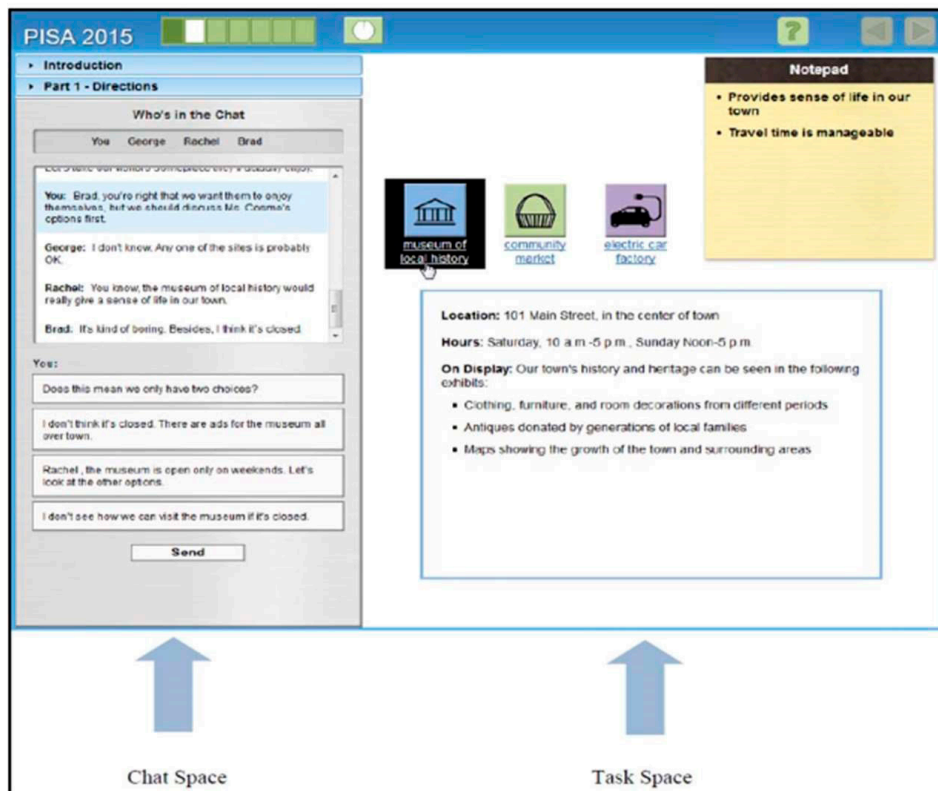


Atriði þar sem próftakar vinna inni í hlut (e. *custom interactive*).

Þessi flokkur atriða á sér ekki hliðstæðu í flokkunarkerfi Scalise og Gifford. Hann inniheldur nýstárlegar gerðir atriða, svo sem hermun (e. *simulation*), fyrir fram samin samskipti við mann-gerving (e. *scripted avatar interactions*), leikjamiðuð (e. *game based*) verkefni og sýndarveruleika (e. *augmented reality, virtual reality*).

Einfaldasta tegund prófatriða í þessum flokki inniheldur hermun og tilraunir, þar sem próftaki framkvæmir tilraun á skjánum og safnar gögnum eða mælingum til að komast að niðurstöðu. Næst koma atriði þar sem próftaki sameinar inntakstæki eins og lyklaborð eða mús við upp-gefin tákni til þess að framkalla svar, t.d. vinna í formúlum, nota reiknivél á skjánum. Svo koma prófatriði sem innihalda samskipti við mann-gervinga, þar sem próftaki leysir atriðið með samskiptum við mann-gerving sem hefur verið forritaður til að gefa ákveðin svör, eftir því hvað próftaki „segir“. Þar næst eru atriði sem byggja á leikjum og samvinnu. Próftaki þarf þá að leysa verkefni sem hefur sömu eiginleika og leikur, t.a.m. markmið, reglur og endurgjöf. Flóknasta tegund atriða í þessum flokki eru atriði sem fela í sér viðbótar- eða sýndarveruleika. Gagnvirkni próftaka og umhverfis gerast í rauntíma þar sem umhverfið veitir margs konar skynupplýsingar og líkir eftir raunverulegum aðstæðum, annað hvort með tilbúnum aðstæðum eða sýndarveruleika (e. *virtual reality* eða *augmented reality*).

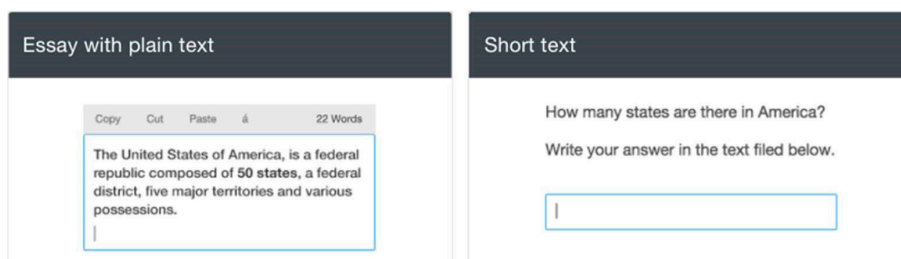
Mynd 6. Dæmi um gagnvirkir atriði (e. custom interactive item) (O’Leary o.fl., 2018).



Atriði þar sem próftakar gera eða útbúa hlut (e. free construction).

Þessi flokkur atriða inniheldur öll atriði þar sem próftakar setja saman eða útbúa svör sjálfir, þ.e. svörin eru ekki valin eða framsett í prófatriðinu sjálfu. Í þessum flokki fela einföldustu atriðin í sér að próftaki skrifi röð stuttra svara, þar sem hvert svar byggir á því sem kemur á undan. Svörin mynda þannig einfalda röksemdafærslu. Næst koma atriði sem krefjast þess að próftaki setji fram lengri textasvör þar sem stikkorð eða tillögur leiða hann áfram í skrifunum. Því næst eru ritgerðarspurningar, þar sem próftaki skrifar lengri texta í skorðuðu eða óskorðuðu svarumhverfi, þ.e. á einum skjá eða þar sem próftaki getur skrunað eða fært sig til í skjámynd. Svo koma atriði þar sem próftaki veitir munnlegt svar við röð spurninga. Flóknust í þessum flokki eru atriði þar sem próftaki teiknar form með frjálsum stíl eða notar gögn sem hann safnar á einhvern hátt sjálfur, til að svara atriðinu.

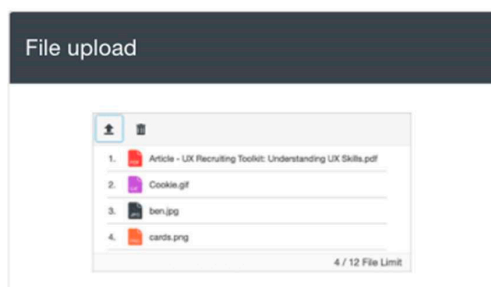
Mynd 7. Dæmi um atriði frá Learnosity þar sem próftakar gera eða útbúa hlut (Murray (e.d.)).



Atriði þar sem próftakar vinna með hlut (e. *upload*).

Þessi flokkur prófatriða krefst þess að próftaki vinni með stærri einingar stafrænna hluta. Hér eru einföldustu tegundir atriða tilraunir eða sannanir, þar sem próftakar framkvæma raunverulega tilraun eða safna gögnum á annan hátt og hlaða svo upp niðurstöðum sínum á rafrænu formi. Næst koma skýrslur, þar sem próftakar skrifa skýrslu eða grein og hlaða upp skjölunum á viðeigandi formi. Svo eru hljóð- eða myndbandsverkefni, þar sem próftaki tekur upp hljóð eða á myndband kynningu eða flutning á ákveðnu efni og hleður upp skránni. Svo eru umræður eða rökræður, þar sem próftaki leysir verkefni sem felur í sér umræður eða rökræður og hleður svo upp hljóð- eða myndbandsupptöku af úrlausninni. Loks eru flóknustu atriðin í þessum flokki atriði sem krefjast þess að próftaki leysi verkefni sem fela í sér einhvers konar kennslu eða þjálfun annarra, taki það upp og hlaði svo upp hljóð- eða myndbandsupptökunni. Hamer og Jongkamp gerðu svo í kerfi sínu ráð fyrir aukaflokki sem nær yfir þau atriði sem passa ekki inn í neinn þessara fyrir fram skilgreindu flokka.

Mynd 8. Atriði frá Learnosity þar sem próftakar hlaða upp hlut (Murray (e.d.)).



Prófakerfi

Helstu prófakerfi sem eru á opnum markaði bjóða upp á margar tegundir TVPA. Fjöldi ólíkra tegunda prófatriða er einn af þeim þáttum þar sem kerfin keppa á, þannig að mikið er lagt upp úr því að hafa sem flestar tegundir prófatriða. Þetta má meðal annars sjá út frá því að í sumum tilvikum virðist einungis blæbrigðamunur á ólíkum tegundum prófatriða. Þetta á einkum við um einfaldari prófatriði, þ.e. þar sem lítill sveigjanleiki er í svarlið atriða. Þegar horft er til flóknari prófatriða geta skil milli stafrænna prófa og hefðbundins námsmats orðið óljós, t.d. þegar stafræn tækni er notuð sem hluti af námsmati sem felst í kynningu á verkefni eða þegar skrifaður texti er meginviðfangsefni námsmats.

Learnosity

Prófkerfi *Learnosity* býður upp á um 60 gerðir TVPA, en mörg önnur prófafyrirtæki notast við það kerfi. Þau TVPA sem *Learnosity* býður upp á eru flokkuð í níu flokka; fjölval (e. *multiple choice*), útfylling (e. *fill in the blanks/cloze*), flokkun, pörun og röðun (e. *classify, match & order*), ritun og upptökur (e. *written and recorded*), auðkenning og teikun (e. *highlight and drawing*), stærðfræðitengd atriði, gröf (e. *graphing*), myndrit (e. *charts*), efnafræðitengd atriði (e. *chemistry*) og loks önnur atriði sem falla ekki í fyrrnefnda flokka. Í þessum flokkum eru svo

mismunandi gerðir spurninga þar sem próftakar nota t.a.m. hreyfiprófatriði (e. *drag-and-drop*) og fellilista (e. *drop-down lists*) við úrlausn.

Inspira

Inspira er kerfi frá Noregi og er notað af Háskóla Íslands og Bifröst. Það er í hraðri þróun þó að það bjóði eins og er ekki upp á mjög margar tegundir prófatriða. *Inspira* er dýrt kerfi og aðgangur að gögnum er nokkuð óljós. Það hafa verið að koma upp vandamál með hugbúnaðarstuðning hjá þeim og tengingar við innskráningarkerfi. Úrval TVPA er takmarkað en hægt að forrita inn í að einhverju leyti sem er þó flókið og kallar á viðhaldsvinnu. Vinnsla í kerfinu er nokkuð auðveld og verð byggir á hverri prófatöku.

TAO

TAO er opinn hugbúnaður en býður upp á hýsingar með mismunandi lausnum sem eru misdýrar. Aukalegur forritunarkostnaður er hár hjá þeim miðað við gögn úr sænsku prófakerfisútbóði. Kerfið hefur verið þýtt að hluta til á íslensku. Viðmótið er þó gamaldags og nokkuð erfitt að vinna í því. Þróun virðist ekki mjög hröð samanborið við mörg önnur fyrirtæki og úrval TVPA er takmarkað. Verð er miðað við hverja prófatöku auk annarra hluta.

Assessment Systems (Ava og FastTest)

Fyrirtækið er aðallega starfandi í Bandaríkjunum og er ekki með starfsstöð í Evrópu. Fyrirtækið er nokkuð stórt á markaði fyrir réttindapróf af ýmsu tagi. Fyrirtækið hefur verið að ganga í gegnum miklar breytingar en allri forritunarvinnu hefur verið úthýst til annarra landa og hafa samskipti við tækniðstoð (e. *support*) verið misgóð.

Ava

Ava er nýtt kerfi frá *Assessment Systems* sem ætlað er að taka við af *FastTest* þegar fram líða stundir. *Ava* byggir á grunni frá *Learnosity* og er með flestar tegundir prófatriða sem nokkurt kerfi býður upp á. Hjá *Ava* er mikill fókus á próffræðilega eiginleika kerfa og aukin áhersla virðist vera á persónumiðaðri próf. Mikil þróun og breytingar eru í gangi í hugbúnaðarmálum hjá þeim.

FastTest

FastTest er eldra kerfi frá *Assessment Systems* og var notað í samræmdu könnunarprófunum á Íslandi 2016–2020. Forritið er enn notað í Arabísku furstadæmunum en búast má við því að notkun verði bráðum hætt. Verð er miðað við hverja prófatöku auk annarra hluta.

Wiseflow/Uniwise

Wiseflow er kerfi frá dönsku fyrirtæki sem byggir í grunninn á Learnosity og er með flestar tegundir prófatriða sem nokkurt kerfi býður upp á. Viðmótið getur verið nokkuð óaðgengilegt en margir módúlar eru í boði fyrir kerfið og þeir eru opnir fyrir viðbótum. Mikil þróun og breytingar eru í gangi í hugbúnaðarmálum hjá þeim. Kerfið er þegar til á færeysku og þýðingarmál eru nokkuð einföld. Gjaldtaka miðast við hvern nemanda óháð prófafjölda.

SoNET

SoNET gerðu samning við Svía um prófakerfi, en kerfið er almennt lokað og erfitt að nálgast upplýsingar. Það er að sameinast öðru fyrirtæki í Bretlandi og er ekki vitað um nákvæma stöðu á þeim samruna. Kerfið sjálft er aðeins komið til ára sinna en er með nokkrar áherslur á einstaklingsmiðuð próf. Svíar hafa aðeins verið að lenda í GDPR-vandamálum þar sem SoNET er staðsett í Ástralíu, auk annarra smávægilegra vandamála. Ekki er vitað hvernig gjaldtöku er háttað.

Tafla 2. *Upplýsingar um framboð TVPA hjá prófafyrirtækjum/prófakerfum sem skoðuð voru við gerð skýrslunnar.*

Tegund TVPA	Digi-exam	Dugga	Inspira	Wise-flow	TAO	Janison	Ada	Fast-Test
MCQ standard	x	x	x	x	x	x	x	x
Multiple response	x	x	x	x	x	x	x	x
Cloze drag and drop		x		x			x	
Cloze drop down				x	x		x	x
Cloze text	x	x	x	x	x	x	x	x
Label image drag and drop				x			x	
Label image drop down				x			x	
Label image text				x			x	
Classification				x	x	x	x	x
Match list		x		x	x	x	x	
Order list		x		x	x	x	x	
Sort list				x	x	x	x	
Essay rich text				x	x	x	x	
Essay plain text		x	x	x	x	x	x	x
Open question	x			x			x	x
Audio recorder			x	x	x	x	x	
Video recorder				x		x	x	
Drawing				x		x	x	
Token highlight		x		x	x		x	
Hotspot				x	x	x	x	
Math		x	x	x	x		x	
Graphing				x			x	
Number line drag and drop				x			x	
Number line plot				x			x	
Line chart				x			x	
Histogram				x			x	
Chemistry				x			x	
Forritun			x					
Fjöldi ólíkra tegunda prófat-riða	4	9	7	27	14	13	27	7

Rannsóknir

Töluverður skortur er á rannsóknum á TVPA og þá sérstaklega rannsóknum á nemendum á grunnskólastigi. Þetta er mjög bagalegt fyrir aðila sem þróa próf og þurfa að velja tegundir prófatriða sem nota skal. Ljóst er að af þeim rannsóknum sem gerðar hafa verið taka flestar til fullorðinna einstaklinga. Þar að auki hefur tæknileg þróun verið hröð síðustu ár og eldri rannsóknir því að mörgu leyti úreltar.

Í nýlegri rannsókn á einbeitingu eða þátttökuvilja próftaka (e. *engagement*) á aldrinum 9–18 ára var hlutfall ágiskana í svörum próftaka við TVPA lægra en í fjölvalsspurningum, sem bendir til þess að TVPA veki meiri áhuga og þátttöku próftaka en hefðbundin atriði (Wise, Soland og Dupray, 2021). Möguleg skýring á því er að TVPA kallar á meiri gagnvirkni en hefðbundin prófatriði. Auk þess innihalda TVPA gjarnan töflur, myndir eða gröf, sem rannsóknir hafa bent til að séu meira áhugahvetjandi (e. *motivating*) en atriði sem innihalda bara texta (Wise, Soland og Dupray, 2021). Hins vegar er á engan hátt ljóst hvort, eða að hvaða marki, þessi aukni áhugi eða virkni tengist færninni sem meta skal eða hvort þetta opni leið fyrir annars konar færni til að hafa áhrif á frammistöðu og um leið niðurstöðu á prófinu. Það er seinna atriðið, að ná til hugsmíða, en ekki það fyrra, að vekja áhuga, sem er lykill að betri mælingu á hugsmíð (Haladyana og Downing, 2004; Kane, 2006; Messick, 1989, 1993, 1996; Newton og Shaw, 2012).

Samantekt á gerðum TVPA sem lögð eru fyrir nemendur í Bandaríkjunum á aldrinum 5–18 ára benti til þess að hreyfiprófatriði (e. *drag-and-drop*) væru algengust. Í greiningunni voru eiginleikar prófatriðanna einnig skoðaðir með tilliti til samsvörunar við kröfur hugsmíðarinnar sem próf metur (e. *construct fidelity*). Meirihluti hreyfiprófatriða hafði takmarkaða samsvörun við kröfur hugsmíðar en aðrar gerðir TVPA höfðu miðlungs eða mikla samsvörun (Russell og Moncaleano, 2019). Höfundar telja skýringuna liggja að hluta til í því að hreyfiprófatriði (e. *drag-and-drop*), sem hafa lága samsvörun, eru í raun atriði sem hægt er að setja fram sem fjölvalsspurningar, t.a.m. hreyfiprófatriði þar sem próftaki á að flokka í annan af tveimur flokkum, en slík atriði eru nokkuð algeng (Russell og Moncaleano, 2019). Höfundar þessarar skýrslu komust að svipaðri niðurstöðu, að margar tegundir tæknivæddra prófatriða væru í reynd flóknar útfærslur á fjölvalsspurningum.

Önnur rannsókn á nokkrum gerðum TVPA benti til þess að TVPA með háa samsvörun við kröfur hugsmíðarinnar (e. *high construct fidelity*) sem próf metur, t.d. hljóðræn atriði, séu betur til þess fallin að meta flókna hugræna færni en TVPA með lægri samsvörun (e. *low construct fidelity*), t.d. fjölval (e. *multiple response*), útfylling (e. *fill in the blanks*) og raðsvörun (e. *ordered response*). Þar af leiðandi ætti ekki að gera ráð fyrir því að öll TVPA geti metið flókna hugræna færni betur en hefðbundin prófatriði eða að innihaldsréttmæti prófa aukist við það eitt að bæta slíkum atriðum við (Qian o.fl., 2017). Auk þess benti þáttagreining Wan og Henly (2012) til þess að þær gerðir TVPA sem voru skoðaðar og fjölvalsspurningar mældu sambærilegar hugsmíðar, þ.e. TVPA og fjölvalsspurningar mynduðu ekki tvo aðskilda þætti heldur einn sameiginlegan þátt.

Samantekt

Hér hefur verið farið yfir helstu eiginleika tölvuvæddra prófatriða (TVPA) og það að hvaða marki þau eru svipuð eða frábrugðin prófatriðum á pappír. Tölvuvædd prófatriði eru byggð upp með sama hætti og prófatriði sem eru á pappír, þau samanstanda af spurnarlið sem leggur viðfangsefni fyrir próftaka og af svarlið þar sem próftaka er gefinn kostur á að leysa viðfangsefni sem lagt er fyrir hann. TVPA eru í öllum meginráttum eins og hefðbundin prófatriði sem lögð eru fyrir á pappír. Rafrænt eða stafrænt umhverfi opnar ákveðna möguleika sem pappír gerir ekki en á móti kemur að próf á pappír hafa einnig sína styrkleika. Markmið þessarar skýrslu er að veita yfirlit um fjölbreytileika tölvuvæddra prófatriða.

Mikilvægt er að kostir TVPA liggi fyrir þegar ákveðið er að beita tiltekinni gerð prófatriða við námsmat. Margar gerðir TVPA eru nauðalíkar prófatriðum sem hönnuð eru fyrir próf sem lögð eru fyrir á pappír. Það er í raun viðfangsefnið sem prófatriði leggur fyrir próftaka sem ræður því hversu krefjandi námsmatið er fyrir próftaka fremur en uppbygging prófatriða.

Eftir að hafa skoðað framboð á TVPA hjá ólíkum prófakerfum og lesið umfjöllun um TVPA í skýrslum og fræðiritum eru niðurstöður höfunda þær að meginþorri TVPA sem boðið er upp á eru í reynd flóknar útgáfur fjölvalsspurninga, rannsóknir á eiginleikum TVPA eru mjög takmarkaðar og að rök sem færð eru fyrir kostum TVPA og notkun þeirra eru alla jafna skoðanir og getugátur höfunda. Fyrsta niðurstaðan, að margar TVPA séu flóknar útgáfur fjölvalsspurninga, eru hvorki rök með eða á móti því að nota TVPA. Að vissu er þetta gamalkunnugt stef úr prófgerðarfræðinni (Haladyana og Rodriguez, 2013; Roid og Haladyana, 1982), að fjölvalsspurningar eru mjög sveigjanlegar og að unnt er að útfæra þær á ólíkan máta eftir aðstæðum (Haladyana og Rodriguez, 2013; Osterlind, 1997). Þetta þýðir að margt af því sem vitað er um fjölvalsspurningar á einnig við um TVPA sem virka í reynd sem fjölvalsspurningar. Til dæmis að það hve miklar hugrænar kröfur prófatriði gerir til próftaka fer ekki eftir formi spurningar heldur því hvernig viðfangsefni er sett fram og hvernig próftaki þarf að vinna með það til að svara prófatriðinu. En einnig virðist sem ábendingar „prófgerðarfræðinnar“ „best practice“ og galla í gerð prófatriða (Haladyana og Downing, 2002; Haladyana og Rodriguez, 2013; Osterlind, 1997) eigi fullt erindi til þeirra sem þróa TVPA eða útbúa prófatriði inn í sniðmát sem aðgengileg eru í notendahugbúnaði. Skoðanir sem settar hafa verið fram sem rök fyrir notkun TVPA lúta meðal annars að því að TVPA auki virkni eða áhuga próftaka og að stafrænt umhverfi TVPA liggi nær veruleika barna og unglinga í dag. Hvoru tveggja eru atriði sem skipta máli og mikilvægt er að huga að við þróun og skipulag á matstækjum (Kingston, Scheuring og Kramer, 2013; Wender og Walker, 2006). Óljóst er hins vegar hvort að TVPA virki áhugahvetjandi eða séu nær veruleika barna og unglinga og sé þannig fullnægjandi ástæða til að nota slík prófatriði nema þau styrki einnig mat á flóknari hugrænni færni (Russell, 2019). Rannsóknir á TVPA eru yfir höfuð takmarkaðar. Lítið hefur verið rannsakað hvort TVPA auðveldi það að meta flóknari hugræna getu nemenda og einnig að hvaða marki þau ná að auka virkni eða þátttöku nemenda í námsmati. Þetta er sláandi í ljósi þess að hundruð rannsókna voru gerðar á kostum og takmörkunum fjölmargra tegunda prófatriða á pappírsformi yfir margra áratuga skeið.

Þótt mikið hafi verið fjallað um mögulega kosti TVPA er skortur á rannsóknum og leiðbeiningum um hvernig eigi að búa til slík atriði (Sireci og Zenisky, 2006; Wan og Henly, 2012). Fjölvalsspurningar eru hins vegar skilvirkar í skorun og uppsetningu, til er mikið af leiðbeiningum um hvernig best sé að hanna slík atriði og með þjálfun er mögulegt að setja upp fjölvalsspurningar sem meta flóknari hugræna færni (Haladyna og Rodriguez, 2013; Osterlind, 1997). Þegar atriði eru valin fyrir próf þarf að hafa í huga hvort og þá hvernig þau bæta prófið. Varað hefur verið við því að nota TVPA einungis vegna þess að þau gefi prófum nýstárlegt og tæknilegt yfirbragð (Parshall o.fl., 2010; Schechinger, 2012).

Heimildir

- Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., og Wittrock, M.C. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives: Abridged edition*. New York. Longman.
- Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. Í M. von Davier, E. Gonzalez, I. Kirsch, K. Yamamoto (Ritstjórar). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 25–45). Springer, Dordrecht.
- Bennett, R. E. (1993). On the meaning of constructed response. Í R. E. Bennett og W. C. Ward (Ritstjórar). *Construction versus choice in cognitive measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Bls. 1–28. Lawrence Earlbaum.
- Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *The Journal of Technology, Learning and Assessment*, 1. <https://ejournals.bc.edu/index.php/jtla/article/view/1667>
- Bennett, R.E., Morley, M. og Quardt, D. (1998). *Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests*. ETS Research Report Series, 1998(2), i-37. <https://doi.org/10.1002/j.2333-8504.1998.tb01794.x>
- Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation*, 22(1). <https://doi.org/10.7275/70yb>
- Cizek, G. J. (2020). *Validity: An Integrated Approach to Test Score Meaning and Use*. Routledge.
- Fizpatrick, R. og Morrison, E. J. (1971). Performance and product evaluation Í R.L. Thorndike (ritstjóri), *Educational measurement* (2. útgáfa, bls. 237–301). WASHINGTON, DC: American Council on Education.
- Haladyna, T. M. og Downing, S. M. (2004). Construct irrelevant variance in high stakes testing. *Educational Measurement: Issues & Practice*, 23(1), 17–27.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5
- Haladyna, T.M. og Rodriguez, M.C. (2013). *Developing and validating test items*. Routledge.
- Hambleton, R.K. (2002). New CBT Technical Issues: Developing Items, Pretesting, Test Security, and Item Exposure. Í C. N. Mills, M. T. Potenza, J. J. Fremer, W. C. Ward (Ritstjórar). *Computer-Based Testing: Building the Foundation for Future Assessments*. Lawrence Earlbaum.

- Hamer, R. og Jongkamp, C. (2021). IB-Cito taxonomy for digital assessment items. [Fyrirlestur]. <https://www.ibo.org/research/assessment-research/presentations/>
- Hochstetter, A., Page, A., Chang, Y. og Cappaert, K. (2018). *Validity inferences for different types of technology-enhanced items*. Presentation at the Annual Meeting of the National Council on Measurement in Education, New York, NY.
- Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15. doi:10.1111/jedm.2003.40.issue-1
- Kane, M. T. (2006). Validation. Í R.L. Brennan (Ritstj.). *Educational measurement*, 4. útgáfa bls. 17–64. American Council on Education/Prager.
- Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50, 1–73. DOI: 10.2307/23353796
- Kingston, N. M., Scheuring S. T., og Kramer, L. B., (2013). Test Development Strategies. Í K. F. Geisinger (ritstjóri). *APA Handbook of Testing and Assessment in Psychology volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*. American Psychological Association, Washington, DC.
- Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4), 212–218.
- Lane, S. og Iwatani, E. (2016). Design og performance assessment in education. Í S. Lane, S.M. Downing & T.M. Haladyna (ritstjórar), *Handbook of test development*. 2. útgáfa (pp. 274–293). New York, NY: Routledge.
- Lane, S. og Stone, C. A. (2006). Performance assessment Í R.L. Brennan (ritstjóri), *Educational measurement* (4. útgáfa, bls. 387–431). Westport, CT: American Psychological Association.
- Measured Progress/ETS Collaborative. (2012). *Smarter balanced assessment consortium: Technology-enhanced items*. Measured Progress
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955–966. <https://doi.org/10.1037/0003-066X.30.10.955>
- Messick, S. (1989). Validity. Í R. L. Linn (ritstjóri) *Educational measurement* (3. útg., bls. 13–103). Phoenix, AZ: American Council on Education, Oryx Press.
- Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. Í R. E. Bennett og W. C. Ward (Ritstjórar). *Construction versus choice in cognitive measurement*. Bls. 75–105. Lawrence Earlbaum.
- Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, 22(2), 13-23. <https://doi.org/10.3102/0013189X023002013>
- Messick, S. (1996). Validity of performance assessment. Í G. W. Philips (Ritstjórar). *Technical issues in large-scale performance assessment*. Bls. 1–18. National Center of Education Statistics.

Millman, J. og Greene, J. (1989). The specification and development of tests of achievement and ability. Í R. L. Linn (ritstj.) *Educational measurement* (3. útg., bls. 335–366). Phoenix, AZ: American Council on Education, Oryx Press.

Murray, R. (e.d.). What are technology-enhanced items? <https://learnosity.com/what-are-technology-enhanced-items/>

Newton, P. E. og Shaw, S. D. (2012). *Validity in educational and psychological assessment*. Cambridge / Sage.

O’Leary, M., Scully, D., Karakolidis, A. og Pisia, V. (2018). The state-of-the-art in digital technology based assessment. *European Journal of Education*, 53(2), 1–16.

Osterlind, S. J. (1997). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance and Other Formats*. Kluwer.

Parshall, C.G. (2002). Item Development and Pretesting in a CBT Environment. Í C. N. Mills, M. T. Potenza, J. J. Fremer, W. C. Ward (Ritstjórar). *Computer-Based Testing: Building the Foundation for Future Assessments*. Lawrence Earlbaum.

Parshall, C.G., Harmes, J.C., Davey, T. og Pashley, P.J. (2010). Innovative items for computerized testing. Í W. Van der Linden, Glas, C. A. W. (Ritstjórar), *Elements of adaptive testing* (pp. 215–230). Springer, New York, NY.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative Item Types for Computerized Testing. Í W. Van der Linden, Glas, C. A. W. (Ritstjórar), *Computerized Adaptive Testing: Theory and Practice* (pp. 129–148). Kluwer Academic Publisher

Perie, M. og Huff, K. (2015). Determining Content and Cognitive Demand for Achievement Tests. Í Lane, S., Raymond, M.R., og Haladyna, T.H. (Ritstj.). *Handbook of Test Development*, 2. útgáfa. New York, NY. Routledge.

Qian, H., Woo, A. og Kim, D. (2017). Exploring the psychometric properties of innovative items in computerized adaptive testing. Í H. Jiao og R.W. Lissitz (ritstjórar), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective*, bls. 97–118.

Rodriguez, M.C. (2002). Choosing an item format. Large-scale assessment programs for all students: Validity, technical adequacy, and implementation, Í G. Tindal, T. M. Haladyna (Ritstjórar). *Large-scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*. Bls. 213–231. Routledge.

Roid, G. H. og Haladyna, T. M. (1982). *A technology for test item writing*. Academic press.

Russell, M. (2006). *Technology and assessment: The tale of two perspectives*. Greenwich, CT: Information Age Publishing.

Russell, M. (2016). A Framework for Examining the Utility of Technology-enhanced Items. *Journal of applied testing technology*, 17(1), 20–32. <http://www.jattjournal.com/index.php/atp/article/view/89189/67798>

- Russell, M. og Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs. *Educational Assessment*, 24(4), 286304.
- Scalise, K. (2012). Using technology to assess hard-to-measure constructs in the common core state standards and to expand accessibility. *ETS invitational research symposium on technology enhanced assessments*, Princeton, NJ. doi:10.1094/PDIS-11-11-0999-PDN
- Scalise, K. og Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, 4(6).
- Scalise, K., Wilson, M. og Gochyyev, P. (2021) A Taxonomy of Critical Dimensions at the Intersection of Learning Analytics and Educational Measurement. *Frontiers in Education*, 2021(6), 656525. doi: 10.3389/educ.2021.656525
- Schechinger, H. (2012). *Innovative computerized test items: A review*. Lawrence, KS: The Center for Educational Testing and Evaluation.
- Schmeiser, C.B. og Welch, C.J. (2006). Test development. Í R.L. Brennan (ritstjóri), *Educational measurement* (4. útgáfa, bls. 307–553). Westport, CT. American Council on Education/Prager.
- Sireci, S.G. og Zenisky, A.L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. Í S.M. Downing & T.M. Haladyna (ritstjórar), *Handbook of test development* (pp. 329–348). New York, NY: Routledge.
- Swiggett, W. D., Kotloff, L., Ezzo, C., Adler, R., Oliveri, M. E. (2014). Usability of Interactive Item Types and Tools Introduced in the New GRE® revised General Test. ETS Research Report No. RR–14-28 doi.org/10.1002/ets2.12028
- Tinkelman, S.N. (1971). Planning the objective test. Í R.L. Thorndike (ritstjóri), *Educational measurement* (2. útgáfa bls. 46–80). Washington, DC: American Council on Education.
- Wan, L. og Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58–78.




MENNTAMÁLASTOFNUN
2156