Inhoud

	Voorwoord	9
Hoofdstuk 1	Constructie van tests, schalen en vragenlijsten	11
1.1	Gebruik van de termen test en schaal	11
1.2	Fasen in een valideringsonderzoek	11
1.3	Unidimensionaliteit	15
Hoofdstuk 2	Uitvoeren en verslag leggen van een factoranalyse	19
2.1	Achtergrond	19
2.2	Leerdoelen van dit hoofdstuk	21
2.3	Definitie van een elementair rapport van een factoranalyse	21
2.4	Doorlopend voorbeeld	22
2.5	Design	24
2.6	Mate van controle	25
2.7	Geaggregeerde data	25
2.8	Hypothesen	28
2.9	Analysemethode	33
2.10	Schatters	38
2.11	Plot van de factorladingen	45
2.12	Toetsing	47
2.13	Beslissing	51
2.14	Interpretatie	55
2.15	Samenvatting elementair rapport	59
2.15.1	Voorbeeld 1	60
2.15.2	Voorbeeld 2	62
2.16	Beknopt rapport	63
2.16.1	Voorbeeld 1	64
2.16.2	Voorbeeld 2	64
2.17	Visualiseren: het lezen van een ladingenplot	64
2.18	Appendix bij hoofdstuk 2	67
2.18.1	De Tucker-Lewis-index en de comparative fit-index	67
2.18.2	De relatie tussen uitkomsten van SPSS en LISREL	68
2.18.3	Het gebruik van de term confirmatieve factoranalyse	69

Hoofdstuk 3	Het vergelijken van meerdere factoranalyses	71
3.1	Achtergrond	71
3.2	Leerdoelen	71
3.3	Wanneer factoranalyses vergelijken?	71
3.4	Het probleem	72
3.5	De basisprincipes	73
3.6	Uitwerking van de basisprincipes	74
3.6.1	Zuinigheid	74
3.6.2	Goodness-of-fit	74
3.6.3	Interpreteerbaarheid	75
3.7	Voorbeelden	77
3.7.1	Vergelijking van meerdere analyses in het onderzoek van Diesfeldt	77
3.7.2	De Big Five, Six, Seven, Eight, Nine, Ten	82
3.8	Computationele problemen bij factoranalyse	84
3.8.1	Sommige varianties 0	84
3.8.2	Geen convergentie	85
3.8.3	Communaliteit groter dan 1	85
3.8.4	Correlatiematrix niet positief definiet	86
3.8.5	Hessiaan niet positief definiet	87
Hoofdstuk 4	Uitvoeren en verslag leggen van een betrouwbaarheids-analyse	89
4.1	Achtergrond	89
4.2	Leerdoelen	89
4.3	Elementair rapport van een betrouwbaarheidsanalyse	90
4.4	Doorlopend voorbeeld	90
4.5	Design	90
4.6	Mate van controle	92
4.7	Geaggregeerde data	92
4.8	Analyse	92
4.9	Schatters	97
4.10	Toetsing	100
4.11	Beslissing	100
4.12	Interpretatie	102
4.13	Samenvatting	108
4.14	Beknopt rapport	109
Hoofdstuk 5	Uitvoeren en verslag leggen van een Rasch-analyse	111
5.1	Achtergrond	111
5.2	Leerdoelen	111
5.3	Het probleem van factoranalyse	112
5.4	Basisconcepten van IRT	115

5.5	Elementair rapport van een Rasch-analyse	116
5.6	Doorlopend voorbeeld	116
5.7	Design	116
5.8	Mate van controle	117
5.9	Hypothese	117
5.10	Geaggregeerde data	120
5.11	Analyse	121
5.12	Schatters	124
5.13	Toetsing	127
5.14	Beslissing	128
5.15	Interpretatie	128
5.16	Samenvatting elementair rapport Rasch-analyse	130
5.17	Beknopt rapport van een Rasch-analyse	132
5.18	Het 3PL-model	132
Hoofdstuk 6	Uitvoeren en verslag leggen van een Mokken-analyse	135
6.1	Achtergrond	135
6.2	Leerdoelen	136
6.3	Elementair rapport van een Mokken-analyse	136
6.4	Doorlopend voorbeeld	136
6.5	Design	136
6.6	Mate van controle	137
6.7	Hypothese	137
6.8	Analyse	139
6.9	Geaggregeerde data	140
6.10	Schatters	143
6.11	Toetsing	144
6.12	Beslissing	148
6.13	Interpretatie	150
6.14	Samenvatting elementair rapport Mokken-analyse	152
6.15	Beknopt rapport van een Mokken-analyse	154
Hoofdstuk 7	Opgaven	155
	Antwoord-format	155
	Antwoordblad factoranalyse	173
	Referenties	175
	Register	189

1 Constructie van tests, schalen en vragenlijsten

1.1 Gebruik van de termen test en schaal

Laten we eerst proberen de termen te stipuleren. Een **meetinstrument** is elke methode die leidt tot kwantitatieve data. Een **test** is een meetinstrument dat bestaat uit meerdere componenten, **items** genoemd, op grond waarvan voor elke persoon een enkele **totaalscore** wordt bepaald. Een **vragenlijst** kan dus ook een test zijn, namelijk als men een totaalscore berekent. Een test hoeft geen *prestatietest* te zijn, maar kan ook een meetinstrument voor attitudes, emoties en sociaal gedrag zijn. Om verwarring te vermijden gebruikt men vaak het woord **schaal** in plaats van test. Zowel de woorden test als schaal worden ook gebruikt voor meetinstrumenten die uit meerdere, samenhangende tests bestaan, die dan **subschalen** of **subtests** worden genoemd. Het woord schaal wordt echter ook gebruikt om een test aan te duiden waarvan de items goed bij elkaar passen. Een schaal die uit subschalen bestaat, zou dan eigenlijk een meetinstrument moeten worden genoemd, terwijl de subschalen schalen zouden moeten heten. Kortom, iedereen roept maar wat. Dus waarom zou ik met die fraaie traditie breken?

1.2 Fasen in een valideringsonderzoek

Bij de constructie van een test moet om te beginnen de validiteit en betrouwbaarheid worden onderzocht. Dit moet natuurlijk worden gedaan voordat de test feitelijk wordt gebruikt. Zo'n onderzoek zullen we hier een **valideringsonderzoek** noemen. In een valideringsonderzoek komen vaak de volgende stappen aan de orde. De hoofdpunten (1, 2, ...) staan in volgorde van tijd, terwijl binnen een hoofdpunt de volgorde van onderpunten (a, b, ...) meestal minder van belang is. Over elk van deze punten kun je een boek schrijven.

1 Voorbereiding

a *Keuze van het soort eigenschappen dat gemeten zal worden.* Voor elke psychologische eigenschap moet in beginsel een aparte subschaal van meerdere items gemaakt worden.

b *Exploratie van het domein met literatuur en interviews*. Stel dat je bijvoorbeeld hebt bedacht dat je een schaal voor agressiviteit wilt hebben. In de literatuur zoek je dan op of zo'n schaal misschien al bestaat, en of de schalen die bestaan, bruikbaar zijn in jouw geval. Het gebruik van al bestaande schalen verhoogt de vergelijkbaarheid van je onderzoek. Stel dat de conclusie is dat je toch een nieuwe schaal moet maken. Dan moet je je toch eerst verdiepen in welke dingen vaak worden gezien als uiting van agressiviteit. Dat fungeert als basis voor het bedenken van de items. Verder kan het geen kwaad om eerst eens te spreken met personen uit de doelgroep, zodat je weet wat er op dat gebied leeft.

2 *Formulering van de items*
Daarbij moet ook steeds beoordeeld worden in hoeverre op inhoudelijke gronden verwacht kan worden dat de items geschikt zijn. Hierbij komen onder andere de volgende facetten aan de orde.

a *Inhoud van de individuele items*. Elk item moet duidelijk passen bij het gekozen domein. Als je een schaal voor agressiviteit maakt, dan moet je er niet een item in stoppen dat behalve van iemands agressiviteit ook nog van andere eigenschappen afhangt, zoals 'als het op mijn werk niet goed gaat, dan maak ik sneller kwetsende opmerkingen tegen mijn partner'. De antwoorden op die vraag zullen namelijk ook afhangen van de vraag of iemand werk en een partner heeft.

b *Representativiteit van de verzameling items*. De items samen moeten een goede vertegenwoordiging van het domein zijn. Bij een schaal voor agressiviteit moet je bijvoorbeeld niet alleen vragen stellen over agressiviteit tijdens het uitgaan, maar ook over school-, werk- en thuissituaties. Merk op dat dit in strijd kan zijn met punt (a).

c *Aantal items*. Per subschaal moet je voldoende items hebben, rekening houdend met het feit dat bij de analyses misschien nog 30% van de items afvalt. Het is moeilijk te zeggen wat moet worden verstaan onder 'genoeg items', maar een schaal van minder dan 10 items zal weinig indruk maken. Bedenk hierbij dat zelfs een tentamen van 40 multiplechoice-items met 4 antwoordcategorieën nog behoorlijk onbetrouwbaar is, in de zin dat iemand met ware score 5 op een schaal van 0 tot 10 met 95% kans een score tussen 3 en 7 krijgt (Ellis, 2004, p. 236). Naarmate de test belangrijker is, en bijvoorbeeld leidt tot belangrijke beslissingen over de persoon, moeten er hogere eisen aan het aantal items worden gesteld. Een praktische beperking is wel dat een te groot aantal items ertoe kan leiden dat de proefpersonen ze niet meer serieus beantwoorden.

d *Precieze formulering van de items*. De items moeten niet voor meerdere interpretaties vatbaar zijn. De formulering moet qua begrijpelijkheid en taalgebruik zijn aangepast aan de doelgroep en het doel van de meting. Een bekend voorbeeld is de regel dat je geen dubbele ontkenningen zou mogen gebruiken. Of dat een goede regel is, hangt ervan af. Als je willekeurige scholieren van het dorpje Groesbeek interviewt over hun ervaringen in de plaatselijke snackbar,

waarschijnlijk wel. Voor een tentamen van rechtenstudenten zou het een idiote regel zijn, want juridische teksten staan bol van de vijfdubbele ontkenningen, en het tentamen dient nu juist te meten of men zulke formuleringen begrijpt.

e *Inhoud en aantal van de antwoordcategorieën*. De antwoordcategorieën moeten zo worden gekozen dat een redelijke spreiding te verwachten is. Items waarop bijna iedereen hetzelfde antwoord geeft, zijn niet informatief. Als je in de Rotterdamse studentenpopulatie vraagt 'hoe vaak ga je naar de kerk', dan zijn de antwoordcategorieën 'minder dan 1 keer per week / 1 keer per week / 2 keer per week of vaker' waarschijnlijk nutteloos, terwijl die in een streng christelijk dorp wel onderscheidend kunnen zijn. Een te klein aantal antwoordcategorieën leidt ertoe dat de antwoorden te weinig informatie bevatten. Meestal wordt aangeraden zo'n 5 tot 7 antwoordcategorieën te gebruiken.

f *Expertoordeel*. In dit stadium kan het ook raadzaam zijn om de mening van een panel van experts te vragen over de items. Bijvoorbeeld: in 2007 was ik betrokken bij de constructie van een vragenlijst om innovatiekracht van zorgconcerns te meten. Er zijn in Nederland allerlei organisaties en adviseurs die zich hebben gespecialiseerd in het stimuleren en begeleiden van innovatie. Zulke mensen zijn al jarenlang de hele dag bezig met innovatie, dus je mag verwachten dat ze een idee hebben van wat het is. Een onderdeel van de constructie was daarom dat er een groep deskundigen op het gebied van innovatie werd uitgenodigd om samen de items te bespreken. De centrale vraag is daarbij of de items samen het begrip 'innovatie' vangen.

3 *Planning van de eerste afname*

De eerste afname zal data opleveren op grond waarvan de schaal nog aangepast kan worden. In de planning moet worden bedacht hoe die data geanalyseerd zullen worden en welke data dan voor de analyses nodig zijn. Daarbij moet aan het volgende worden gedacht.

a *Gebruik van meerdere beoordelaars*. Als er gebruik wordt gemaakt van observatoren of beoordelaars, dan moet ook de interbeoordelaarsbetrouwbaarheid worden bepaald en daarvoor is het nodig dat er meerdere beoordelaars voor dezelfde proefpersoon worden gebruikt.

b *Andere variabelen die gemeten moeten worden*. Dat zijn bijvoorbeeld achtergrondvariabelen en verwante meetinstrumenten.

c *Aantal proefpersonen dat nodig is*. Het aantal proefpersonen dat nodig is, hangt mede af van de statistische eigenschappen van de data die verkregen worden (MacCallum, Browne & Sugawara, 1996). Toegepast op schaalconstructie zouden er minstens 100 personen nodig zijn. Barrett (2007) stelt echter dat elk artikel met een structural equation model (het soort model dat ook in schaalconstructie wordt gebruikt) bij minder dan 200 proefpersonen sowieso verworpen moet worden. Wirth en Edwards (2007) suggereren ook dat minimaal 200 proefpersonen nodig zijn. Voor sommige analyses zijn echter minstens 1000 personen nodig (Flora & Curran, 2004).

4 *Eerste afname van de schaal*
Hierbij worden de data verzameld die in de volgende stappen geanalyseerd zullen worden.

5 *Analyse van data van individuele items*
Dit is eigenlijk een soort voorwas of voorselectie, waarbij de slechtste items alvast verwijderd worden.
   a *Interbeoordelaarsbetrouwbaarheid van items*. Als de items zijn gebaseerd op observatie of beoordelingen, moet worden vastgesteld dat verschillende observatoren een hoge mate van overeenstemming hebben. Want als men het over de elementaire data al niet eens is, dan kun je wel ophouden met dat item. De overeenstemming wordt meestal bepaald met Cohen's kappa of met een intraclass-correlatie.
   b *Variantie van de items*. Items met een kleine variantie zijn meestal minder geschikt omdat zij weinig bijdragen aan het onderscheiden van personen. Als ik bijvoorbeeld de tentamenvraag stel 'hoeveel is 1 + 1?', dan heeft iedereen dat goed (variantie 0) en dan kan ik de vraag dus net zo goed niet stellen. En een vraag die iedereen fout beantwoordt, is evenmin informatief. Items met een kleine variantie zullen ook vaak een kleine of zelfs negatieve bijdrage aan de betrouwbaarheid hebben. Als grens wordt wel geopperd dat voor items die in gehele getallen worden gescoord, de variantie minstens 1 moet zijn. Dat is echter geen harde grens, en geen enkele grens is goed verdedigbaar. Mijn advies is om items te verwijderen als hun variantie nul is, en daarnaast op grond van factoranalyse, IRT-analyse of betrouwbaarheidsanalyse, maar niet op grond van hun variantie.
   c *Scheefheid en eentoppigheid van de verdeling van de items*. Grote verschillen in de scheefheid van items hebben invloed op de correlaties, en daarmee op de uitkomsten van de factoranalyse. Bij de meeste vormen van factoranalyse wordt verondersteld dat de items normaal verdeeld zijn. Aangezien items meestal een klein aantal discrete antwoordcategorieën hebben (bijvoorbeeld 0-1-2-3-4), kunnen zij niet normaal verdeeld zijn. Niettemin is het nuttig om de afwijking zo klein mogelijk te houden door items te verwijderen die sterk afwijken van symmetrie en eentoppigheid. (Eentoppig betekent dat het histogram eruitziet als een klok en niet als een badkuip.)

6 *Analyse van de relaties tussen de items*
Hierbij is de vraag of de items hetzelfde meten. Dat wordt ook wel *unidimensionaliteit* of *homogeniteit* genoemd. Dit is van belang om te rechtvaardigen dat de itemscores worden samengevat tot een enkele totaalscore per persoon.
   a *Correlaties tussen de items*. Items van dezelfde schaal moeten positief correleren (waarbij je wel items mag spiegelen op grond van hun inhoud). Dit, omdat zij in essentie dezelfde eigenschap moeten meten.

b *Factoranalyse van de items*. Hierbij wordt nauwkeuriger onderzocht of de correlaties tussen de items het rechtvaardigen om te geloven dat de items hetzelfde meten. Als blijkt dat de items niet unidimensioneel zijn, dan moet de schaal misschien worden gesplitst in subschalen of moeten er items worden verwijderd.
c *Analyse van de interneconsistentie-betrouwbaarheid*. Hierbij wordt onderzocht of de schaal genoeg items heeft om de somscore te zien als een betrouwbare meting. Hoeveel items daarvoor nodig zijn, hangt ook af van de hoogte van de correlaties tussen de items. Verder worden items met een negatieve bijdrage aan de betrouwbaarheid uit de schaal verwijderd.

7 *Normering*
Hierbij onderzoek je wat voor de te onderscheiden doelgroepen de gemiddelden, standaarddeviaties en percentielen zijn. Dit, om aan te geven wat 'normaal' is.

8 *Analyse van de relaties van de testscores met andere variabelen*
Dit komt vaak pas aan de orde bij latere afnamen van de test, en is een doorlopend proces van het uitbreiden van informatie over de test.
a *Test-hertest-betrouwbaarheid*. Hierbij wordt onderzocht hoe stabiel de scores in de loop der tijd zijn. Het is niet altijd nodig dat de scores stabiel zijn; dat hangt af van het construct en de theorieën daarover. Bijvoorbeeld: een test die de gemoedstoestand van die dag beoogt te meten, hoeft volgend jaar niet dezelfde scores op te leveren. Maar het is meestal wel nuttig dat de stabiliteit bekend is.
b *Criteriumvalidering*. Hierbij wordt onderzocht in hoeverre de test andere variabelen kan voorspellen. Hierbij gaat het erom het praktische gebruik van de test te onderbouwen.
c *Constructvalidering*. Hierbij wordt onderzocht of de test de relaties heeft die theoretisch verwacht worden. Hierbij gaat het vooral om de theoretische interpreteerbaarheid van de test.

In de volgende hoofdstukken wordt alleen ingegaan op punt 6. Een centraal begrip daarin is 'unidimensionaliteit'. Dit zal nu worden geïntroduceerd.

1.3 Unidimensionaliteit

Eenvoudig gezegd wil **unidimensionaliteit** zeggen dat de items hetzelfde meten. Dit kan worden gezien als een onderdeel van constructvaliditeit. Constructvaliditeit wil zeggen dat de items het beoogde construct meten, en daarvoor is op zijn minst nodig dat ze allemaal hetzelfde meten. Het soort data dat wordt gebruikt is bij unidimensionaliteit echter meer vergelijkbaar met interneconsistentie-betrouwbaarheid. Unidimensionaliteit heeft ook invloed op die betrouwbaarheid.

Zoals in de vorige paragraaf al is opgemerkt, is unidimensionaliteit van belang om te rechtvaardigen dat de itemscores worden samengevat tot een enkele totaalscore per persoon. Dat verdient enige uitleg. Het lijkt misschien volkomen vanzelfsprekend om itemscores bij elkaar op te tellen. Maar hoe weet je dan wélke items je kunt optellen? Mag je depressie-items ook bij intelligentie-items optellen? Nee, natuurlijk niet. Net zomin als appels en peren. Maar hoe weet je dan dat je niet met alle bestaande tests appels en peren aan het optellen bent? Jij kunt als testconstructeur wel vinden dat de items mooi bij elkaar passen, maar zeggen de data dat ook? Blijf je ook nog denken dat twee items hetzelfde meten als ze een negatieve correlatie blijken te hebben?

Met een test probeer je menselijk gedrag te kwantificeren, en het is helemaal niet vanzelfsprekend dat dit mogelijk is. Veel mensen zijn diep beledigd als je hun mooie gevoelens in die nare getallen wilt persen. En ze hebben gelijk, want je kunt wel van alles bij elkaar optellen, maar hoe kom je erbij dat dat goed is? Het is waar dat al een eeuw lang bij allerlei tests de scores worden opgeteld, maar dat bewijst niet dat dit ook verstandig is.

De cruciale vraag is in hoeverre de somscore een goede samenvatting geeft van het gedrag dat de persoon op de test vertoonde. Met andere woorden, in hoeverre de itemscores adequaat worden samengevat. Stel bijvoorbeeld dat een vragenlijst uit 50 ja/nee-vragen bestaat, die we coderen met 0 (nee) en 1 (ja). Iemand met een somscore van 25 heeft dus 25 keer ja gezegd. Maar zonder nadere informatie weten we dan nog niet op wélke vragen de persoon ja heeft gezegd. Het kunnen net zo goed de eerste 25 vragen als de laatste 25 vragen zijn. Of alleen de oneven vragen. Het is dus denkbaar dat in een groep personen met dezelfde totaalscore twee subgroepen bestaan die een totaal tegenovergesteld gedrag vertonen: waar de ene groep ja zegt, zegt de andere nee. In zo'n geval is de somscore geen goede samenvatting van de itemscores. Als de betreffende vragenlijst over agressiviteit ging, dan zou dit betekenen dat er niet één soort agressiviteit bestaat, maar minstens twee verschillende soorten, bijvoorbeeld directe en indirecte agressiviteit. De consequentie daarvan is dat men niet één testscore per persoon moet gebruiken, maar minstens twee verschillende testscores. Onderzoek naar unidimensionaliteit dwingt psychologen dus om hun concepten (in dit geval agressiviteit) te differentiëren (in dit geval in directe agressiviteit en indirecte agressiviteit) als de data daar aanleiding toe geven. Zo wordt voorkomen dat al te gemakkelijk dingen bij elkaar worden gevoegd die eigenlijk verschillend zijn.

In het bovenstaande werd een voorbeeld gegeven waarbij de somscore geen goede samenvatting was van de itemscores. Wanneer is de somscore wel een goede samenvatting van de itemscores? Dat is als personen met dezelfde somscore in de regel ongeveer dezelfde itemscores hebben, afgezien van random ruis. De modellen die gebruikt worden om unidimensionaliteit te onderzoeken, beschrijven dit nauwkeurig. Deze modellen hebben allemaal de volgende aannamen:

1. *Unidimensionaliteit*. Elke persoon kan worden gekarakteriseerd met één enkel getal dat aangeeft in welke mate die persoon de eigenschap heeft die men wil meten. Dit getal is in principe niet bekend. Het wordt daarom de latente trek

genoemd. Bij een intelligentietest zou dit iemands onbekende echte intelligentie zijn, en bij een depressieschaal is het iemands onbekende echte depressiviteit. Het betreffende getal wordt meestal aangeduid met θ (de Griekse letter theta), maar in sommige contexten ook met τ (de Griekse letter tau, van 'true score').

2. *Monotoniciteit*. De verwachte score op een item neemt toe met θ. Als de score op item i wordt aangegeven met X_i en de verwachte waarde met E (van expectation), dan geldt dus dat

   $$E(X_i \mid \theta) = f_i(\theta)$$

   waarbij f_i een stijgende functie is.

3. *Lokale onafhankelijkheid*. Binnen een groep subjecten met dezelfde waarde van θ hangen de items niet samen. In de totale populatie mogen er wel hoge correlaties tussen items zijn.

Bijvoorbeeld: iemand die een lage agressiviteit heeft, moet *op alle gebruikte items* een lage score hebben (afgezien van ruis), en als de agressiviteit van die persoon groter wordt, dan moet dit op alle items naar voren komen. Een ander voorbeeld is een tentamen. Daarbij is het wenselijk dat goede studenten op alle vragen een betere kans hebben. Een tentamenvraag waarop de goede studenten juist slechter scoren, dus waarop je minder punten krijgt naarmate je de stof beter kent, is onwenselijk.

Unidimensionaliteit en monotoniciteit zijn empirisch niet te onderscheiden. Als je alleen maar unidimensionaliteit aanneemt, zonder monotoniciteit, dan is het niet toetsbaar. Datzelfde geldt voor lokale onafhankelijkheid. Daarom wordt de term unidimensionaliteit in de praktijk vaak gebruikt voor deze drie assumpties samen. Het woord heeft dus twee betekenissen.

Hierboven werd gesuggereerd dat unidimensionaliteit, monotoniciteit, en lokale onafhankelijkheid samen toetsbaar zijn. Hoe dan? Wel, een eenvoudige voorspelling die volgt uit de genoemde assumpties is: alle items van de schaal moeten niet-negatieve correlaties met elkaar hebben (Mokken, 1971; Mokken & Lewis, 1982; Holland & Rosenbaum, 1986; Ellis & Junker, 1997; Junker & Ellis, 1997; Junker & Sijtsma, 2001b).

Wat is het verschil tussen unidimensionaliteit en interneconsistentie-betrouwbaarheid? Unidimensionaliteit zegt dat de items hetzelfde meten afgezien van ruis, maar zegt niets over de grootte van de ruiscomponent. Interneconsistentie-betrouwbaarheid zegt iets over de grootte van de ruiscomponent in de totaalscore, maar zegt niets over de vraag of de items hetzelfde meten. Bijvoorbeeld: de rekenitems '3 + 4 = ?' en '5 + 2 = ?' zijn unidimensioneel, maar hun totaalscore is niet betrouwbaar omdat het slechts twee items zijn. Daarentegen heeft de totaalscore van een IQ-test meestal een hoge betrouwbaarheid, maar intelligentie is niet unidimensioneel omdat er verschillende soorten intelligentie zijn (bijvoorbeeld fluid intelligence en crystallized intelligence).