



Psychometrische kanttekeningen bij de onderzoeken van Bosman en Schraven

Cito heeft in de periode 2006-2010 de LOVS-toetsen Spelling voor groep 3 tot en met 8 uitgegeven. In deze toetsen komen zowel open opgaven (in de vorm van een dictee) als meerkeuzeopgaven voor. Hiermee wordt naast actieve spelling ook het kunnen herkennen van spelfouten als onderdeel van de spellingvaardigheid meegenomen. De LOVS-toetsen Spelling hebben twee doelen: niveaubepaling en progressiebepaling. Gezien de beoordeling van de COTAN dekken de toetsen deze doelen. Bosman en Schraven (Schraven, Bosman & van Eekhout, 2010; Bosman & Schraven, 2013) menen dat het toetsen van spelling door meerkeuzeopgaven psychometrisch onjuist is en tot didactische problemen leidt. Deze conclusies worden echter gebaseerd op basis van onderzoek dat een aantal aantoonbare onjuistheden en belangrijke omissies bevat. De belangrijkste hiervan worden hieronder weergegeven.

De uitspraak dat de LOVS-toetsen Spelling M4 onbetrouwbaar zijn, is onjuist.

- De uitspraak is gedaan op een zeer kleine, homogene groep, en betreft dus niet de betrouwbaarheid voor de gehele populatie waarvoor de toets bedoeld is.
- Als wij al mee willen gaan met de eis dat voor deze homogene populatie de toets betrouwbaar moet zijn, dan zien wij dat Bosman en Schraven niet de gehele toets bespreken, maar twee keer de halve: de module (subtoets) met dictee-opgaven met alpha .74 en de meerkeuzemodule (subtoets) met alpha .75. Volgens de normen van de COTAN is de betrouwbaarheid voor subtoetsen voldoende als deze boven de .70 ligt (Evers, Lucassen, Sijtsma & Meijer, 2010; p.33).
- De twee modules worden dan wel afzonderlijk afgenomen, maar geen van de modules leidt tot een afzonderlijke vaardigheidsscore. Alleen de scores *samen* worden door Cito als volledige spellingtoets gepresenteerd, en leveren een vaardigheidsschatting op. Zelfs in de homogene subgroep van Bosman en Schraven heeft deze toets dan een betrouwbaarheid van .84, wat door COTAN als voldoende wordt bestempeld voor toetsen op basis waarvan belangrijke beslissingen genomen worden. Voor voortgangstoetsen zoals deze LOVS-toets kan dit zelfs als goed gezien worden.
- In de reguliere populatie (waarvoor de betrouwbaarheid feitelijk berekend moet worden) heeft de toets een nog hogere betrouwbaarheid. Daarom is de toets op dit punt ook als goed beoordeeld door COTAN.

Aangezien bovenstaande uitspraak onjuist is, zou het wetenschappelijk redelijk zijn als deze geschrapt zou worden.

De uitspraken gebaseerd op de eigen gemaakte toetsen met dictee-opgaven zijn onjuist.

- Bosman en Schraven geven aan dat **alle** door hunzelf gemaakte subtoetsen met 25 opgaven in hun steekproef een alpha hebben van lager dan .70, met een gemiddelde betrouwbaarheid lager dan .50. Volgens de normen van de COTAN zijn de subtoetsen daarmee onvoldoende betrouwbaar om scores van individuen goed te bepalen.
- Ondanks dat de subtoetsen niet betrouwbaar zijn, worden toch individuele scores vergeleken, waarop ook de conclusies gebaseerd worden. Dit gebeurt met correlaties en met verschijscores. De verschijscores die gerapporteerd zijn, zijn echter nog onbetrouwbaarder dan de subtoetsen zelf (zie voor berekeningen van de betrouwbaarheid van verschijscores de formules in Van den Brink en Mellenbergh, 1998, p.82). In geval van het onderzoek van

- Bosman en Schraven is de betrouwbaarheid van de gerapporteerde verschillscore .50. Aanzienlijk lager dan nodig is voor het doen van verantwoorde uitspraken.
- De enige score die wel voldoende betrouwbaar is voor het vergelijken van scores van individuen, is de totaalscore op de verzameling van alle 100 door Bosman en Schraven gemaakte dictee-opgaven. Deze toets heeft een betrouwbaarheid van .79. Hoewel deze toets twee keer zo lang is als de gehele LOVS-toets Spelling M4, is deze nog altijd minder betrouwbaar dan de LOVS-toets Spelling M4 voor de betere spellers in deze populatie ($\alpha = .84$).
 - Deze totaalscore laat een correlatie met de meerkeuzemodule zien van .68. Dit is het enige resultaat dat over deze score gerapporteerd wordt.
 - De correlatie na attenuatiecorrectie – wat een validiteitsmaat is (Van den Brink en Mellenbergh, 1998, p.82) – is .88. Dit valt hoog te noemen en kan als bevestiging gezien worden van de conclusie dat meerkeuzeopgaven en dictee-opgaven (vrijwel) hetzelfde meten en op één schaal te plaatsen zijn.
 - Schraven (2011) geeft op pagina 4 van zijn stuk aan dat de betrouwbaarheid van een toets een noodzakelijke voorwaarde is voor validiteit: “Wanneer een toets bij iedere afname een heel andere score oplevert (dus onbetrouwbaar is), dan is het zeker dat deze het beoogde kenmerk niet goed meet (dus niet valide is).” Doordat de subtoetsen met 25 opgaven van Bosman en Schraven onbetrouwbaar zijn volgens de normen van de COTAN, zijn deze toetsen volgens de redenering van Schraven dus ook niet valide. De zelfgemaakte subtoetsen met 25 opgaven zijn dan ook onbruikbaar en er kunnen geen conclusies uit getrokken worden.

Het zou wetenschappelijk redelijk zijn de resultaten op de ondeugdelijke toetsen van Bosman en Schraven te schrappen. Alleen het resultaat van hun toets met 100 dictee-opgaven is aantoonbaar voldoende betrouwbaar. Als we onze conclusies baseren op deze score dan zien we dat de dictee-opgaven en meerkeuzeopgaven vrijwel hetzelfde meten.

Door de omissie van de resultaten van de correlaties voor toetsen met dictee-opgaven onderling wordt een onjuist beeld geschetst over het effect van vraagvorm.

- Alleen de correlaties tussen de subtoetsen met dictee-opgaven en de meerkeuzemodule zijn gegeven. De correlaties tussen de subtoetsen met dictee-opgaven onderling ontbreken.
- De correlaties tussen de subtoetsen met dictee-opgaven onderling zijn echter vergelijkbaar in grootte met de geschatte correlaties tussen subtoetsen met de dictee-opgaven en de meerkeuzemodule (zie van Berkel, Keuning & Tomesen, 2013; voor berekening zie Bijlage 1).
- De samenhang tussen een meerkeuze-subtoets en een subtoets met dictee-opgaven is dus vergelijkbaar met de samenhang tussen twee subtoetsen met dictee-opgaven.
- De voorspelling van een toets met meerkeuzeopgaven naar een andere toets met dictee-opgaven is zodoende ook niet slechter dan de voorspelling van een toets met dictee-opgaven naar een andere toets met dictee-opgaven.
- Er kan dus niet gezegd worden dat toetsen met dictee-opgaven onderling wel hetzelfde meten, maar dat toetsen met dictee-opgaven en meerkeuzetoetsen iets anders meten.
- De vraagvorm heeft dus bijzonder weinig invloed op de verschillen in de correlatie.
- De lage correlaties in het onderzoek van Bosman en Schraven zijn grotendeels te wijten aan de onbetrouwbaarheid van de door de auteurs zelf gemaakte toetsen. Bij de betrouwbare toetsen in het onderzoek ($\alpha > .70$) zijn de correlaties heel redelijk. De correlaties na attenuatiecorrectie zijn zelfs goed te noemen.

Het zou wetenschappelijk correct zijn als de auteurs op basis hiervan erkennen dat de vraagvorm relatief weinig effect heeft op de correlaties tussen de verschillende toetsen.

Literatuur

Berkel, S. van, Keuning, J. & Tomesen, M. (2013). Het onderzoek van Bosman en Schraven in een ander perspectief. Een reactie op het artikel 'Cito-spellingtoets schaadt het spellingonderwijs'. *Basisschoolmanagement*, februari, 10-12.

Bosman, A. & Schraven, J. (2013). Cito-spellingtoets schaadt het spellingonderwijs. *Basisschoolmanagement*, februari.

Bosman, A.M.T., Schraven, J.L.M. & van Eekhout, T. (2010). De Cito-spellingtoets: onze bezwaren nader toegelicht. *Orthopedagogiek: Onderzoek en Praktijk*, 49,412-421.

Evers, A., Lucassen, W., Sijsma, K., & Meijer, R.R. (2010). COTAN beoordelingssysteem voor de kwaliteit van test (gewijzigde herdruk). *NIP/COTAN*, Amsterdam..

Schraven, B. (2011). *De omvang van een onderzoek en de validiteit van de Cito-Spellingtoetsen*. www.annabosman.eu.

Schraven, J.L.M., Bosman, A.M.T. & van Eekhout, T. (2010). De nieuwe Cito-spellingtoets ter discussie. *Tijdschrift voor Orthopedagogiek (O en A)*, 49,75-86.

Van den Brink, W.P. & Mellenbergh, G.J. (1998). *Testleer en toetsconstructie*. Boom, Amsterdam.

Bijlage 1: De correlatie tussen de dictees met A-, B-, C- en D-woorden onderling

De correlaties tussen deze subtoetsen worden niet gegeven. Onder de aanname dat de dictees onderling allen even hoog correleren, zijn deze echter wel te achterhalen op basis van gegevens uit Tabel 1 uit Schraven, Bosman en van Eekhout (2010). In die tabel wordt de standaarddeviatie (SD) gegeven van het gemiddelde van de SOM (de som van A, B, C en D): $SD(GEM[A \text{ t/m } D]) = SD((SOM)/4)$, waaruit de SD en de variantie (VAR) van de som A-D af te leiden is: $SD(SOM) = 4 * SD(GEM[A-D]) = 4 * 1,8 = 7,2$ en $VAR(A+B+C+D) = 51,84$

De relatie tussen de variantie van de som (X+Y) en de losse onderdelen (X en Y) is ook bekend:

$$VAR(X+Y) = VAR(X) + VAR(Y) + 2COV(X,Y)$$

Voor 4 onderdelen geldt:

$$VAR(A+B+C+D) = VAR(A) + VAR(B) + VAR(C) + VAR(D) + 2COV(A,B) + 2COV(A,C) + 2COV(A,D) + 2COV(B,C) + 2COV(B,D) + 2COV(C,D)$$

Voor de losse subtoetsen A tot en met D weten we:

open sets	SD(.)	VAR(.)
A	2,9	8,41
B	2,2	4,84
C	1,6	2,56
D	2,2	4,84

Dat betekent dus dat: $51,84 = 8,41 + 4,84 + 2,56 + 4,84 + 2COV(A,B) + 2COV(A,C) + 2COV(A,D) + 2COV(B,C) + 2COV(B,D) + 2COV(C,D)$

Hieruit valt af te leiden:

$$\begin{aligned} 2COV(A,B) + 2COV(A,C) + 2COV(A,D) + 2COV(B,C) + 2COV(B,D) + 2COV(C,D) &= 31,19 \Leftrightarrow \\ 2 * [COV(A,B) + COV(A,C) + COV(A,D) + COV(B,C) + COV(B,D) + COV(C,D)] &= 31,19 \Leftrightarrow \\ COV(A,B) + COV(A,C) + COV(A,D) + COV(B,C) + COV(B,D) + COV(C,D) &= 15,595 \end{aligned}$$

$Correlatie(X,Y) = COV(X,Y) / [SD(X) * SD(Y)]$, dus $COV(X,Y) = Correlatie(X,Y) * SD(X) * SD(Y)$

In dit geval dus :

$COV(A,B) =$	$Corr(A,B) *$	$SD(A) * SD(B) =$	$Corr(A,B) *$	6,38
$COV(A,C) =$	$Corr(A,B) *$	$SD(A) * SD(C) =$	$Corr(A,C) *$	4,64
$COV(A,D) =$	$Corr(A,B) *$	$SD(A) * SD(D) =$	$Corr(A,D) *$	6,38
$COV(B,C) =$	$Corr(A,B) *$	$SD(B) * SD(C) =$	$Corr(B,C) *$	3,52
$COV(B,D) =$	$Corr(A,B) *$	$SD(B) * SD(D) =$	$Corr(B,D) *$	4,84
$COV(C,D) =$	$Corr(A,B) *$	$SD(C) * SD(D) =$	$Corr(C,D) *$	3,52

Dit levert 1 vergelijking met 6 onbekenden op. Als we echter aannemen dat correlaties tussen de verschillende subtoetsen met dicteeopgaven onderling niet verschillen, dan wordt dit een vergelijking met 1 onbekende: $Corr(A,B) = Corr(A,C) = Corr(A,D) = Corr(B,C) = Corr(B,D) = Corr(C,D) = \text{correlatie}$.

$COV(A,B) =$	$Corr(A,B) *$	$SD(A) *$	$SD(B) =$	$Correlatie *$	6,38
$COV(A,C) =$	$Corr(A,B) *$	$SD(A) *$	$SD(C) =$	$Correlatie *$	4,64
$COV(A,D) =$	$Corr(A,B) *$	$SD(A) *$	$SD(D) =$	$Correlatie *$	6,38

COV(B,C) =	Corr(A,B) *	SD(B) *	SD(C) =	Correlatie *	3,52
COV(B,D) =	Corr(A,B) *	SD(B) *	SD(D) =	Correlatie *	4,84
COV(C,D) =	Corr(A,B) *	SD(C) *	SD(D) =	Correlatie *	3,52
SOM =				Correlatie *	29,28

$$\text{Correlatie} = \text{Som}/29,28 = 15,595/29,28 = 0,53$$

Vergelijken we dit met de gemiddelde correlatie van de A tot en met D dictees en de meerkeuzemodule, dan zien we dat die 0,51 is, hetgeen dus amper scheelt met de correlatie tussen de dictees onderling.

De aanname dat de correlatie tussen de dictees onderling gelijk zijn, kan fout zijn, maar stel dat we nu vergelijkbare correlaties invullen voor de dictees onderling als voor de dictees met de meerkeuze-opgaven dan kan dit het volgende opleveren.

COV(A,B) =	Corr(A,B) *	SD(A) *	SD(B) =	0,64 *	6,38 =	4,1
COV(A,C) =	Corr(A,B) *	SD(A) *	SD(C) =	0,37 *	4,64 =	1,7
COV(A,D) =	Corr(A,B) *	SD(A) *	SD(D) =	0,64 *	6,38 =	4,1
COV(B,C) =	Corr(A,B) *	SD(B) *	SD(C) =	0,36 *	3,52 =	1,3
COV(B,D) =	Corr(A,B) *	SD(B) *	SD(D) =	0,65 *	4,84 =	3,1
COV(C,D) =	Corr(A,B) *	SD(C) *	SD(D) =	0,38 *	3,52 =	1,3
SOM =						15,63

$$\begin{aligned} \text{VAR}(A+B+C+D) &= 20,65 + 2 * 15,634 = 51,92 \\ \text{SD}(A+B+C+D) &= 7,21 \\ \text{SD}(\text{gem}(A-D)) &= 1,80 \end{aligned}$$

Oftewel, de correlaties tussen de dicteetoetsen onderling zijn mogelijk exact dezelfde correlaties als tussen dicteetoetsen en meerkeuzetoetsen. Oftewel: een lage correlatie ligt niet aan de vraagvorm, maar aan iets anders.