Bevindingen Expertgroep tussentijdse check IEP 2019, psychometrisch gedeelte.

IDnummer	19.010
Naam Toets	IEP
Aanvrager	Bureau ICE
Datum beoordeling	23 september 2019
Datum definitief rapport Expertgroep	9 oktober 2019

Deze rapportage is gebaseerd op de Verantwoording IEP Eindtoets | Afname 2019 en normeringsonderzoek 2020 (Culemborg: Bureau ICE - juli 2019).

Algemene indruk is dat de rapportage van goede kwaliteit, en dat het normeringsonderzoek adequaat is. Verder lijkt de IEP in de operationele fase in 2019 goed gefunctioneerd te hebben, maar informatie over de betrouwbaarheid ontbreekt. Verder zijn er m.b.t. de IRT analyses nog wel wat vragen over de identificatie van het model en de simulatiestudie die om opheldering vragen.

4 Steekproef

- blz.10. Terecht wordt opgemerkt dat de functie van de pretest met het invoeren van het gezamenlijke anker verandert is en dat de nadruk vooral is komen te liggen op de kwaliteit van nieuw geconstrueerde items ten behoeve van de samenstelling van zowel het gezamenlijk anker als van de IEP eindtoets 2020.
- blz.10. Tekst: “Uitgangspunt is dat ieder te pretesten item door ongeveer 1.000 leerlingen gemaakt wordt, waarmee ruimschoots aan de eis voor steekproefgrootte van 400 observaties per item, conform Meijer et al.(2016) , voldaan wordt.”. Het aantal van 1000 leerlingen wordt in de pretest goed benaderd. Hierbij tekent de Expertgroep echter wel aan dat ze zich niet kan vinden in het door de COTAN genoemde minimum aantal van 400 leerlingen voor het 1PLM. Het is weliswaar zo dat het 1PLM voldoende nauwkeurig te schatten is met dit aantal, maar voor het evalueren van de modelpassing moet ook het meest voor de hand liggende alternatieve model, het 2PLM schatbaar zijn, ook al schat je dat model niet expliciet. Daarvoor zijn ongeveer 1000 leerlingen per item ruim voldoende. Voor de pretest van items voor het gezamenlijke anker (Boekje E) is dit minder belangrijk, omdat bijna alle aanbieders deze items pretesten, zodat voldoende observaties beschikbaar zijn.
- blz.11/12. het rekruteren van de steekproef is duidelijk beschreven en adequaat uitgevoerd. Tabel 4.1 is voldoende informatief. Statistische toetsing van de hypothese van een random steekproef uit een populatie is niet uitgevoerd en niet strikt nodig. Bij grote steekproeven is de toets toch altijd significant door de grote power van de toets. Men zou eventueel een effectgrootte op kunnen nemen, bijvoorbeeld, Cramer’s V of Phi. Die laatste maat is voor de goede orde even uiteengezet in de Appendix A.

5 Design gegevensverzameling

- Het design van de gegevensverzameling is duidelijk beschreven en het design is adequaat.

6 Kalibratie en kwaliteit van de items: IRT-analyse

- blz.23. In de inleiding van dit hoofdstuk wordt aangegeven dat er, om de kwaliteit van de items te bepalen, voor alle drie de onderdelen een IRT-analyse is uitgevoerd waarbij de gegevens van de pretest in 2019 en de IEP Eindtoets 2019 samen zijn geschaald.
- blz. 23. Tekst “Om de voorlopige cesuren van de IEP Eindtoets 2020 te kunnen schatten, is er een simulatie gedaan met 60.000 virtuele leerlingen”. Deze zin is een beetje verwarrend door het gebruik van het woord cesuur. De (voorlopige) cesuren op de latente schaal liggen al vast, maar het zal gaan om de cesuren in termen van de aantal-goed schaal. De (zeer informatieve) tabellen 7.1 en 7.2 wijzen ook in die richting.
- Er zijn DIF analyses uitgevoerd waarvan de resultaten in Bijlage 2 zijn opgenomen. Het gaat om DIF tussen de boekjescombinaties. In de tekst zou iets meer uitleg gegeven kunnen worden ver wat nu precies het doel is van deze DIF analyse. De tabellen en de resultaten zien er goed uit.
- blz. 25. De analyse van de vaardigheidsverdeling van de populatie leiden tot vragen. Het IRT model is geïdentificeerd door de vaardigheidsverdeling van de eindtoetspopulatie 2019 als standaard-normaal te definiëren (zie tabellen 6.1a,b,c). Echter, de analyses zijn uitgevoerd door de pretest in 2019 en de IEP Eindtoets 2019 samen te schalen, en die schaal is al geïdentificeerd door de 2018/2019 ankeritems. Dus die restrictie op de populatieparameters lijkt niet nodig.
- De methode van simuleren voor de tabellen 7.1a,b,c verdient wat uitleg. De eerste stap is waarschijnlijk dat er 60000 theta’s getrokken zijn uit een verdeling, om daarna antwoordpatronen te genereren. Is er daarna via WML bij iedere totaalscore een theta terug geschat? Of is er een andere methode gebruikt. Ook Tabel 7.2 verdient enige uitleg over hoe theta aan een totaalscore gekoppeld is.

8 Functioneren als operationele eindtoets

- Voor de overigens erg informatieve Tabel 8.1, geldt hetzelfde als hetgeen hierboven is opgemerkt bij Tabel 4.1: effectgrootten zouden eventueel behulpzaam zijn. hetzelfde geldt ook voor de verdere gerapporteerde kruistabellen.
- Tabel 8.7 is erg informatief. De IEP zit wat hoger dan het schooladvies. De consequenties hiervan moeten aan de orde komen bij een inventarisatie van de normeringsmethoden van de andere aanbieders, en speciaal het CvTE.
- Opmerking: er wordt geen informatie gegeven over de globale en lokale betrouwbaarheid van de toetsonderdelen en toetsadviezen. De globale en lokale betrouwbaarheid van de onderdelen is rechtstreeks uit de output van de IRT analyse te halen. Maar TIA’s zoals die in Bijlage 5 voor de pretest worden gegeven zijn voor de globale betrouwbaarheid ook adequaat. Voor de globale betrouwbaarheid van de toetsadviezen is de covariantie of correlatiematrix tussen de onderdelen nodig. De methode wordt uiteen gezet in Appendix B.

APPENDIX A: Phi als effectgrootte bij een kruistabel.

Vaak is statistische toetsing van de hypothese van een random steekproef uit een populatie niet informatief. Bij grote steekproeven is de toets toch altijd significant door de grote power van de toets. Een effectgrootte is nuttiger. Phi kan als effectmaat (met waarden Phi ≤ 0.1 klein effect volgens Cohen) berekend worden op basis van de chi-kwadraat toetsingsgrootheid. Hieronder een klein voorbeeldje

Tabel 4.5 Aantal en percentage leerlingen in de populatie en de steekproef naar schooltype

	Populatie	Steekproef			
Stratum	**%**	**M7**	**%**	**E7**	**%**
0-10%	67,3	3095	68,6	1613	68,6
10-25%	22,2	1001	22,2	517	22,0
25-40%	6,6	264	5,9	144	6,1
>40%	3,8	150	3,3	76	3,2

M7 χ^2 (3, N = 4510) = 8,260; p = 0,041; phi = 0,043
E7 χ^2 (3, N = 2350) = 3,743; p = 0,291; phi = 0,040

$$phi = \sqrt{\frac{X^2}{N}}$$

APPENDIX B: Reliability of composite scores.

The composite reliability is a function of the reliability of the individual components, the weights that are assigned to the components the variances and error variances of the component scores, and the covariances between the different components. All this can be combined into the following formula (1):

$$r = 1 - \frac{\sum_{i=1}^{n}(w_i^2 \sigma_{e,Xi}^2)}{\sum_{i=1}^{n}(w_i^2 \sigma_{Xi}^2) + \sum_{i=1}^{n}\sum_{j(\neq i)=1}^{n}(w_i w_j \sigma_{Xi,Xj})} \qquad [1]$$

Where:

w_i is the weight of component *i*;

I. σ_{Xi}^2 is the observed variance of component *i*;

II. $\sigma_{e,Xi}^2$ is the error variance of component *i*, which is a function of the component reliability and the observed variance: $\sigma_{e,Xi}^2$ = (1 - reliability of component i) * σ_{Xi}^2 ;

III. $\sigma_{Xi,Xj}$ is the covariance between the two components, which can be rewritten as: $Correlation_{Xi,Xj} * Standard\ deviation_{Xi} * Standard\ deviation_{Xj}$

Note that the nominator of the ratio in formula 1 is the error variance of the index, while the denominator is the total variance. The error variances of joint count components were calculated from their observed variance and reliability levels. However, two significant problems arose during component reliability computations.

Voorbeeld

Van theta-schattingen (Zowel EAP als WML) kan je, bijvoorbeeld in SPSS of in Excel een Variantie-Covariantie matrix maken. Dat leidt bijvoorbeeld tot de volgende matrix.

Variantie-Covariantie matrix Taal

	Leesvaardigheid	Luisteren	Taalverzorging	Begrippenlijst	Woordenschat
Leesvaardigheid	0.622				
Luisteren	0.279	0.554			
Taalverzorging	0.228	0.190	0.537		
Begrippenlijst	0.431	0.373	0.419	1.376	
Woordenschat	0.352	0.298	0.258	0.460	0.983

Als we de formule (1) toepassen leidt dit tot een betrouwbaarheid van 0.8728