

Kickstart Your Soft Skills

Ontwikkeling en validering vragenlijst

Ontwikkeling in batches

De ontwikkeling van de vragenlijst is verlopen in twee batches. Hier beschrijven we het verloop van de eerste batch. De tweede batch is op een gelijkaardige wijze verlopen. Werken met twee batches gaf ons de mogelijkheid om de werkwijze te testen bij batch 1 en indien nodig aan te passen voor batch 2.

Tabel 1: overzicht van soft skills per batch

Cluster	Batch 1	Batch 2
Interactief	Communiceren	Inleving
Probleemoplossend	Samenwerken	Digitale vaardigheden
Informatieverwerkend	Diversiteit	Klantgerichtheid
Persoonsgebonden	Resultaatgerichtheid	Analyseren
	Flexibiliteit	Kritisch denken
	Creativiteit	Leerbereidheid
	Plannen & organiseren	Zelfreflectie
		Verantwoordelijkheid
		Zelfstandigheid

De selectie van skills voor batch 1 werd gemaakt in het kader van de volgende opportuniteit. In het kader van duaal leren voerde de faculteit Productontwikkeling van de Universiteit Antwerpen een project uit rond soft skills waarbij ze een tool ontwikkelden die ingezet wordt als evaluatie-instrument bij stages. Zeven soft skills uit hun project kwamen overeen met skills uit het KYSS-model. Ze waren zeer gemotiveerd om gebruik te maken van de gevalideerde KYSS-vragenlijst en we kwamen tot een overeenkomst om de zeven gemeenschappelijke soft skills in batch 1 te ontwikkelen.

Hieronder leggen we uit hoe het verloop bij de ontwikkeling van zo'n batch liep en wat de eventuele methodologische wijzigingen zijn tussen de eerste en de tweede batch.

Opstellen van rubrieken per soft skill

De eerste stap in het ontwikkelen van de vragenlijst bestond eruit om voor elke soft skill een rubriek van gedragscriteria met kwalitatieve niveaoverschillen op te stellen. Deze rubriek zou later ingezet kunnen worden als coachingsmateriaal. Voor het opstellen van de rubrieken werden eerst reeds bestaande gevalideerde vragenlijsten en rubrieken verzameld. We baseerden ons hiervoor op wetenschappelijke bronnen, overheidsbronnen en bestaande rubrieken van AP Hogeschool¹. Op basis van deze bronnen werden 20 tot 40 gedragscriteria gedefinieerd, verdeeld over drie niveaus ('beginner', 'gemiddeld' en 'gevorderd'). Voor batch 2 hebben we het aantal gedragscriteria gereduceerd tot 12 à 17 criteria per skill na feedback van de inhoudelijke experts om zo de items makkelijker te kunnen sorteren (zie volgende paragraaf). Bij batch 2 werd dus een strengere selectie gemaakt van bestaande gedragscriteria.

Inhouds- en taalkundige validiteit rubrieken

Vervolgens werd de inhoudsvaliditeit van de gedragscriteria nagegaan door deze voor te leggen aan enkele inhoudelijke experts, komende zowel uit het onderwijs (AP Hogeschool, UA) als uit het werkveld (VDAB, Groep INTRO, Bayer, J-TEC, VOKA). De rubrieken werden geprogrammeerd in een gerandomiseerde sorteeroefening. De experts werden gevraagd om de criteria te sorteren van 'makkelijk' (of items die een beginner moet kunnen) naar 'moeilijk' (of items die een gevorderde moet kunnen). Hierna werd deze feedback als volgt geanalyseerd.

- Wanneer items systematisch door verschillende experts op hetzelfde niveau geschat werden ($SD < 1$), werd het item behouden.
- Van zodra er te veel discrepantie was tussen de meningen van de expert, werd het item weerhouden ($SD > 1$ of minstens 2x niet geselecteerd).
- Ook werd de interbeoordelaar-betrouwbaarheid in acht genomen. Hiervoor werd de item-rest correlatie (RIR) berekend tussen de experts om te bepalen of bv. één bepaalde expert een heel afwijkende mening had. In dat geval zou overwogen worden om deze mening weg te laten. Dit was echter niet het geval.
- Tot slot werd de open feedback van de experts bekeken. Ze konden bij de oefening aangeven of bepaalde criteria niet bij de skill hoorden of er nog bepaalde criteria misten.

Op basis van deze analyses werd een selectie gemaakt van ongeveer tien criteria per skill.

De taalkundige validiteit van deze criteria werd vervolgens nagegaan door een lector Nederlands van AP Hogeschool en een taalcoach bij Groep INTRO en VDAB (samen met enkele werkzoekenden). Ze gingen na of de criteria taalkundig correct zijn en begrijpbaar voor de beide doelgroepen. In het kader van genderanalyse heeft een expert van RoSa vzw (Kenniscentrum voor gender en feminisme) onze vragenlijst beoordeeld en goedgekeurd.

Na het analyseren van de taalkundige feedback, merkten we een spanningsveld tussen het taalniveau van studenten en dat van werkzoekenden bij VDAB en Groep INTRO. Dit spanningsveld hebben we

¹ Zie 'Bijlagen – Tabel 4: Bronnen ontwikkeling rubrieken'

enerzijds voorgelegd aan onze klankbordgroep en anderzijds besproken met Dr. Eva Kyndt, gespecialiseerd in het opstellen en valideren van vragenlijsten. De klankbordgroep gaf ons waardevolle feedback waardoor we het taalniveau van onze vragenlijst konden aanscherpen. Dr. Eva Kyndt adviseerde om de analyses van de pilootstudies eerst voor de doelgroepen apart te doen. Indien de vragenlijst gevalideerd kan worden voor beide doelgroepen, kunnen ze samen geanalyseerd worden en kan bekeken worden of we één vragenlijst kunnen ontwikkelen voor beide taalniveaus (universal design).

Pilootstudie

De wetenschappelijke validiteit van de vragenlijst werd vervolgens nagegaan door een pilootstudie. De pilootstudie voor batch 1 liep van 20 april tot 15 juni en werd uitgezet bij laatstejaarsstudenten uit de professionele bachelors van AP Hogeschool en werkzoekenden bij VDAB en Groep INTRO. Voor beide doelgroepen mikten we op een respons van 200 respondenten. Voor de studenten werd de vragenlijst uitgestuurd via de online leeromgeving van AP Hogeschool. Wanneer een respondent de vragenlijst invulde kreeg deze automatisch een feedbackrapport via de online leeromgeving. Voor de werkzoekenden werd de vragenlijst uitgestuurd via Limesurvey en Qualtrics. Respondenten die een feedbackrapport wensten, konden hun e-mailadres opgeven en kregen daar het rapport toegestuurd. Van zodra we voldoende respons behaalden, werd de vragenlijst afgesloten en de data geanalyseerd.

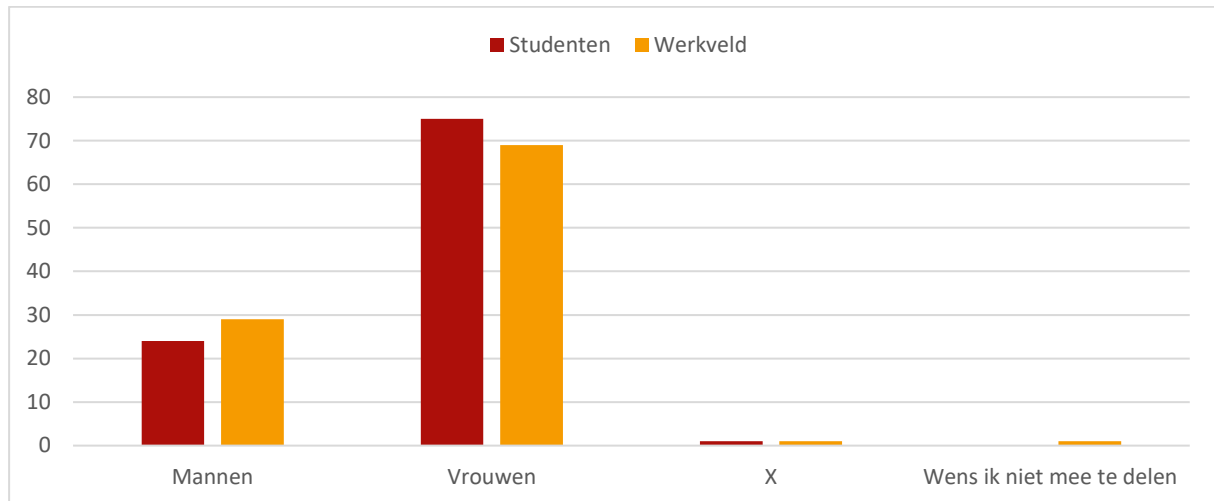
Hieronder volgt een beknopte toelichting van de resultaten van de analyses. Het volledige rapport van de data-analyses van batch 1 en 2 zal toegevoegd worden aan de eindrapportering.

Steekproef

De uiteindelijke steekproef voor batch 1 bedraagt 242 studenten en 141 werkzoekenden. Gezien de beperkte steekproef, zullen extra analyses uitgevoerd worden op de data die in de toekomst via de KYSS-tool verzameld wordt.

De onderstaande figuur toont verder aan dat bij beide doelgroepen de spreiding over geslacht gelijkaardig is. Dit wil zeggen dat bij de analyses de data tussen beide doelgroepen vergeleken mag worden op basis van de gendervariabele.

Figuur 2: Spreiding percentage respondenten over geslacht



Exploratieve en confirmatorische factoranalyses

We gebruikten exploratieve factoranalyse met een oblimin rotatie om na te gaan of we voor elke schaal voldoende kwaliteitsvolle items konden identificeren. Kwaliteitsvol betekende in dit geval een factorlading van 0.50 of hoger. Daarbij trachtten we meteen om de vragenpool te reduceren tot een kleinere verzameling van de beste 6 tot 8 items per schaal, met het oog op de hanteerbaarheid van de uiteindelijke vragenlijst. De uitkomsten van de twee bevragingcontexten (studenten en werkzoekenden) werden vergeleken om in te schatten of één vragenlijst voor de twee contexten een realistische optie was.

Vervolgens gingen we aan de hand van confirmatorische factoranalyses na of de items van de gereduceerde schalen telkens één en hetzelfde construct per schaal meten en of deze schalen betrouwbaar genoeg zijn. Voor alle zeven soft skills hebben we uiteindelijk een schaal kunnen creëren met +- 6 items met een aanvaardbare tot goede constructvaliditeit en betrouwbaarheid.

Hierna bekeken we de samenhang van de verschillende schalen binnen en tussen de vier clusters op basis van de correlaties. De schalen binnen de clusters hangen goed samen, maar er zijn enkele schalen die over de clusters heen een grote samenhang hebben. Van zodra we de analyses van batch 2 uitvoeren, zullen we de samenhang van alle 16 schalen bekijken. Indien we een zeer grote samenhang tussen enkele schalen uit verschillende clusters zien, bekijken we of we de schalen van clusters wisselen. Ook zullen we de correlaties van de vraagitems onderling tussen verschillende schalen bekijken. Wanneer we voor bepaalde items een zeer hoge samenhang zien, zullen we overwegen om die items van schaal te wisselen.

Stabiliteit over doelgroepen en geslacht

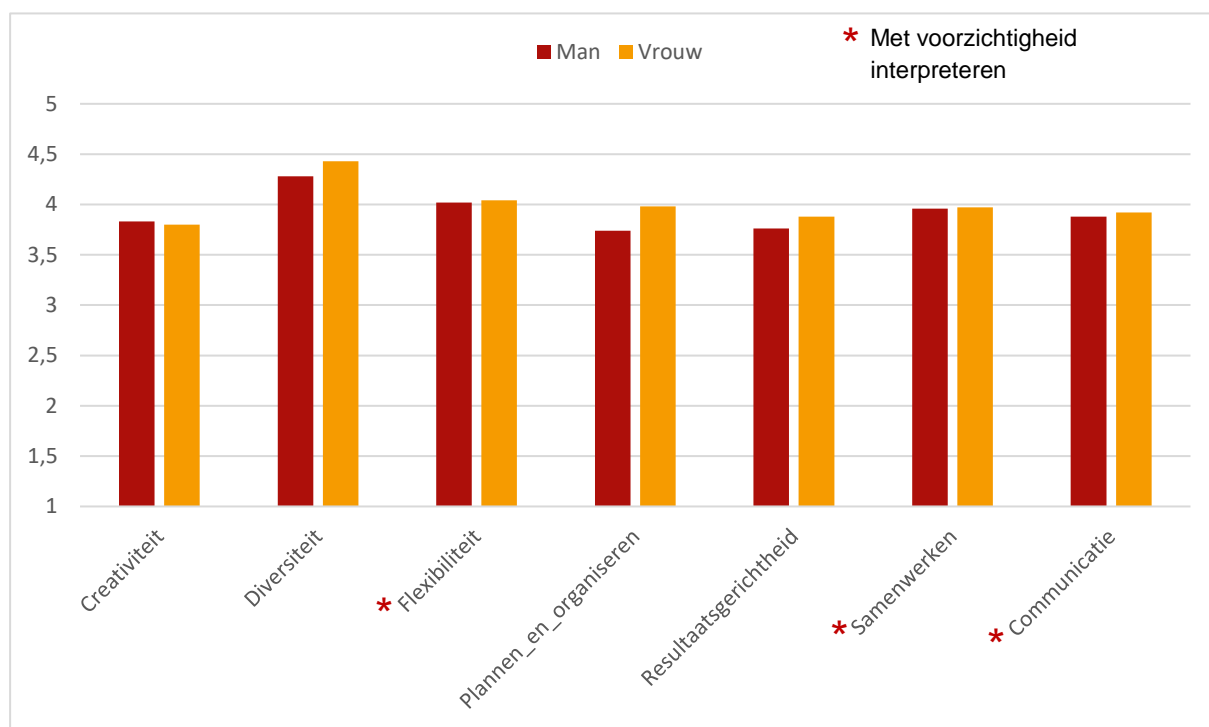
Via stabiliteitsanalyses gingen we na of de schalen stabiel zijn over de doelgroepen en geslacht. Indien wel, zou dit betekenen dat studenten en werkzoekenden, en mannen en vrouwen de schalen op dezelfde manier interpreteren, en bijgevolg zouden we de schaalscores mogen vergelijken op basis van de doelgroep- en/of de gendervariabele.

Uit de stabiliteitsanalyses bleek dat we de schaalscores mogen vergelijken over geslacht, zij het met voorzichtigheid voor 'flexibiliteit', 'samenwerken' en 'communiceren'. Het kan zijn dat mannen en vrouw deze drie schalen anders interpreteren. De schalen waren minder stabiel over de doelgroepen heen, dus schaalscores vergelijken tussen studenten en werkzoekenden wordt best vermeden. Er is weliswaar geen aparte vragenlijst nodig voor beide doelgroepen, maar men moet er wel rekening mee houden dat de vraagitems een verschillende rol spelen bij studenten dan bij werkzoekenden.

We zien hier echter geen probleem in, aangezien het niet onze intentie is om in de toekomst data te vergelijken tussen studenten en werkzoekenden. Het is wel interessant dat lectoren, werkgevers, coaches, enzovoort de data kunnen vergelijken tussen de mannen en vrouwen in hun groep. Zo kunnen ze bijvoorbeeld inspelen op bepaalde noden die heersen binnen de groep. Van zodra we meer data verzamelen binnen de KYSS-tool, plannen we om verder onderzoek te doen naar de minder stabiele schalen voor geslacht. Door de kleine steekproef wegen verschillen over geslacht namelijk zwaarder door. Van zodra we een grotere steekproef hebben, gaan we na of deze verschillen blijven en of we dan onze vraagitems moeten aanpassen of herformuleren.

Wanneer we uiteindelijk de schaalscores vergeleken over geslacht zien we slechts twee significante verschillen. Vrouwen scoorden significant hoger op de schalen 'diversiteit' en 'plannen en organiseren'. Het gaat weliswaar telkens om kleine effecten (2% verklaarde variantie). Over de doelgroepen heen zien we dat enkel de schaalscores voor 'resultaatgerichtheid' significant verschillen (5% verklaarde variantie), maar zoals hierboven vermeld hebben we niet de intentie om studenten en werkzoekenden te vergelijken.

Figuur 3: Spreiding schaalscores over geslacht

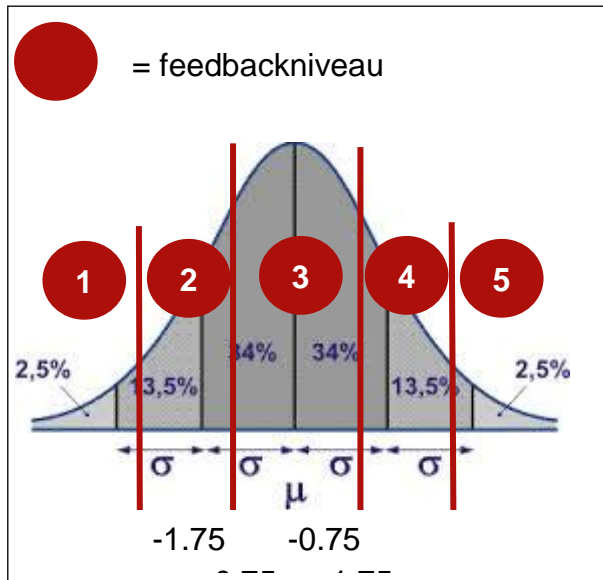


Normering

De normering zal de grensscores bepalen aan de hand waarop we de schaalscores van respondenten onderverdelen en score-specifieke feedback gaan geven. We gaan werken met een statistische normering, wat inhoudt dat we deze baseren op de data die we verzamelen. Momenteel hebben we een kleine steekproef en gaan we uit van een zeer initiële normering, welke we jaarlijks zullen bijstellen op basis van de verzamelde data.

We gaan een normering maken met vijf niveaus waarbij we uitgaan van een normaalverdeling: de grensscore van het middelste niveau staat gelijk aan het gemiddelde, de grensscores voor de andere niveaus staan gelijk aan het gemiddelde min of plus 0.75 en 1.75 standaarddeviatie.

Figuur 4: Voorbeeld normering o.b.v. normaalverdeling



Momenteel zien we dat de normering eerder problematisch is vanwege hoge gemiddeldes en spreiding binnen de schalen. Dit is enerzijds te verklaren vanwege de kleine steekproef, anderzijds zijn de respondenten die vrijwillig meededen aan de pilootstudie ‘gemotiveerde’ respondenten en gaan ze zichzelf mogelijks hoger inschatten. Nog een andere verklaring voor de hoge zelfinschatting zou kunnen zijn dat respondenten tijdens het invullen van de vragen niet voldoende reflecteren over hun werkcontext. In de vragenlijst van batch 1 stond enkel in de inleiding vermeld dat ze de werkcontext in hun achterhoofd moesten houden. Tijdens batch 2 zullen we een contextvraag toevoegen om dit op te vangen. Zoals hierboven vermeld, zullen we in vervolgonderzoek de normering bijstellen op basis van de grotere steekproef.

Batch 2

Voor de negen skills van batch 2 zijn ondertussen ook rubrieken opgesteld. Zoals hierboven vermeld hebben we voor deze skills een strengere selectie gemaakt van criteria uit bestaande gevalideerde vragenlijsten en rubrieken. Elke rubriek bevat tussen de 12 en 17 gedragscriteria. Om de inhouds- en taalkundige validiteit na te gaan, hebben we dezelfde stappen gevolgd als bij batch 1. Doordat we de methodologie voor het ontwikkelen van de vragenlijst al konden uittesten bij batch 1, ging het ontwikkelen van batch 2 opvallend sneller. Tijdens de tweede helft van augustus werden de rubrieken opgesteld en eind september hadden we alle inhoudelijke en taalkundige feedback ontvangen en verwerkt. De pilootstudie zal van start gaan op 1 oktober en we verwachten de data te analyseren in november – december.