Supervisor competence as an assessor of medical trainees

Summary

Chapter 1

Supervisor suitability, competency and qualification to identify the clinical and professional achievement of medical trainees to a required standard logically would seem an essential requirement for training and assessment programmes. Indeed, a *sine qua non* for running a safe and effective healthcare system in which training of new practitioners is an integral part. Any assessor providing educational instruction to medical trainees would be expected to measure with a high degree of precision and accuracy the level of achievement for the required learning exhibited by medical trainees. However this expectation is not supported by much of the assessment literature related to judgement-based assessments. This is of concern given the evidence of a substantive amount of poor and unsafe healthcare as a global problem, in economically developed countries and more frequently in and low-and middle-income countries. The consequences of errors, failures to use effective care, overuse of ineffective care, and disregard of peoples' values and resources for people's health are substantial. The associated waste of equipment, supplies, time, and human spirit causes unacceptable economic costs to healthcare systems. Each and every one of these well-documented failures is directly related to medical practitioners' clinical and professional competencies and medical and scientific knowledge, whether medical error, and overuse or underuse of evidence-based quality medical practice. Concerns about patient safety, geographic variation in patient care unrelated to medical science, and poor "customer service" for patients, have coalesced to call into question the competence of physicians and health care systems in general. The adequacy of the training and assessment of medical trainees to prepare them as clinicians for immediate and long-term clinical practice must be questioned. Even though some evidence exists that enhanced clinical supervision and assessment of trainees can be associated with improved patient-related and education-related outcomes, this evidence is not substantial. This state of affairs is understandable if the training provided is of variable quality and the measurement of the achievement of required clinical and professional standards is inaccurate and imprecise. As a result adequately measuring the quality of an improvement intervention and the outcomes of any educational interventions will be doomed to a continuation of the current state of uncertainty. A scientific approach is needed to overcome instructional and assessment problems arising from such issues as

using assessments without observing trainees in real-life situations, not incorporating multiple perspectives such as peers and patients, and not using measures that are sufficiently accurate to predict clinical outcomes. The minimisation of incompetence in clinical practice is an important simple universal goal for clinical assessment methods. A critical requirement for a scientific approach requires reproducible methodology and measurement. To achieve a scientific level of supervised training and assessment in the workplace, the presence of the required competencies for clinical and professional practice must not only be precisely and accurately measured, but also measurable. The simple aims for the work of this dissertation was firstly to evaluate the need for disruptive change of the *status quo* using a critical appraisal of the validity of supervisor assessments with the results positioned in the context of current knowledge and using fine-grained reliability and validity measures. A second aim was to attempt innovative change with an exploratory approach to validity evidence as a quality improvement method combined with traditional quality control improvement process. The work for the dissertation that constitutes academic complexity is firstly simplification and normalisation of what are considered complex statistical and data-analytic methods. Secondly to use those methods as outcomes, not just as validity evidence but also as quality indicators of judgement-based assessments, and quality indicators of assessors' ability to make those judgements of competence in others.

Chapter 2

In line with the first aim, this chapter critically evaluates the reliability of supervisor assessments and compares that reliability with a commonly used workplace-based assessment, namely the mini-Clinical Evaluation Exercise (mini-CEX). The supervisor assessment is summative and the research uses empirical data. The mini-CEX is meant to be formative and instructional, and the method used to evaluate the reliability were results of an unpublished systematic review. The design and method for the empirical reliability study was a trainee nested in the assessment using Generalisability theory. Score variation attributable to the trainee for each competency-item assessed was estimated by the minimum-norm quadratic unbiased estimator, and that score variance was used to estimate the number of assessments for an adequate reliability value of 0.80. The same measures were obtained from all available reliability studies for the mini-Clinical Evaluation Exercise (mini-CEX) at the time. The trainee score variance for each competency item varied between 5.4% for emergency skills to 37.2% for communication skills, with an average for all competency items of 19.3%; the "Overall rating" competency item was 27.6%. These variance components translated into 70, 7, 17 and 10 assessments needed for a reliability of 0.80, respectively. Importantly, the measures were slightly better than those observed for the mini-CEX for the similar competencies assessed and published in the literature. Supervisors' assessments may be more reliable than assessments using the min-CEX

when comparing trainee score variation and sample number requirements. Both types of assessment provide results that have a substantive level of unreliability. The results support the statement that despite the wide-spread use of rater-based judgements of clinical competence, they are context sensitive and vary between individuals and institutions. To deal adequately with rater-judgement unreliability, evaluating the reliability of workplace rater-based assessments in the local context remains essential. Disruptive change of the status quo is that supervisors' assessments could be used as part of a trainee workplace assessment programme providing sufficient numbers of assessment are performed. All training programmes using supervisors' assessments need to provide that information to the accrediting body as routine evidence of adequacy of the assessment process.

Chapter 3

The study for this Chapter measured construct validity of supervisor competency assessment results, measured the reliability of the identified competency-domain constructs, measured construct stability across subgroups and positioned the results in the established literature. The construct identification was by traditional exploratory factor analysis (EFA), reliability was established using Generalisability theory, factor stability was explored by EFA of subgroups, and the results were compared with previous similar studies. The unit of analysis was each assessment and included all available assessments. Reliability of identified constructs was by variance components analysis of the summed trainee scores for each factor and the number of assessments needed to provide an acceptably reliable assessment using the construct; the reliability unit of analysis was the score for each factor for every assessment. The EFA resulted in 3 factors accounting for only 68 % of the variance with factor 1 having features of a ''general professional job performance'' competency; factor 2 ''clinical skills''; and factor 3 ''professional and personal'' competency. The trainee score variance for the summed competency item scores for the factors were 40.4%, 27.4% and 22.9% respectively. The number of assessments needed to give a reliability coefficient of 0.80 was 6, 11 and 13 respectively. The factor structure remained stable for subgroups of female trainees, Australian graduate trainees, the central hospital, surgeons, staff-specialists, visiting medical officers and the separation into single years. Physicians as supervisors, male trainees, and male supervisors all had a different grouping of items within 3 factors which all had different dimensionality of the competency items indicating the presence of factor instability in some subgroups. Supervisors as a group are assessing a dominant construct domain which is similar to a general professional job performance competency. However, factor structure instability between different populations of supervisors and trainees means that subpopulations of trainees may be assessed differently to others. A lack of competency criterion standardisation of supervisors' assessments brings into question the validity of this assessment method as currently used. These results provide further evidence for a need for improvement and

change, particularly since the results reflect decades of similar historical concern, as history seems to be continually repeating itself but not invoking improvement.

Chapter 4

The information resulting from the previous Chapters have confirmed the variability in the alignment of competency constructs between different assessors and different contexts. Different supervisors and assessors often have contrasting views about competency constructs as well as other latent behavioural constructs, making assessment results very unreliable for what are high stakes assessments. Construct alignment, accuracy and precision for the interpretation of the competency constructs used for trainee assessment is essential so further examination of the validity evidence was undertaken using an evaluation of the internal validity and reliability of competency constructs from supervisors' end-of-term summative assessments for prevocational medical trainees. CFA was used to evaluate assessment internal construct validity. A previously identified hypothesised competency construct model was tested, with comparisons between competing models of potential competency constructs including the competency construct model of the original assessment. The hypothesised competency constructs of "general professional job performance", "clinical skills" and "professional abilities" provided a good model-fit to the data, and a better fit than all alternative models. Model fit indices were $\chi 2/df = 2.8$; RMSEA = 0.073 (CI 0.057-0.088); CFI = 0.93; TLI = 0.95; SRMR = 0.039; WRMR = 0.93; AIC = 3879; and BIC = 4018). The optimal model had adequate measurement invariance with nested analysis of important population subgroups supporting the presence of full metric invariance. Reliability estimates for the competency construct "general professional job performance" indicated a resource efficient and reliable assessment for such a construct (6 assessments for an $R > 0.80$). Item homogeneity was good (Cronbach's alpha = 0.899). Other competency constructs are resource intensive requiring ≥11 assessments for a reliable assessment score. Although internal validity and reliability of clinical competence assessments using judgement-based methods are acceptable when actual competency constructs used by assessors are adequately identified. However, still supervisors are constructing their own individual assessment process independent of others' understanding and what regulatory bodies require, resulting in major unreliability and also common method bias, making the assessment results meaningless.

Chapter 5

The validity of assessment results can be considered a measure of the quality of that assessment process. For judgement-based assessments, construct validity is also a potential measure of assessor ability to provide valid and reliable feedback. The quality of high-

stakes supervisor assessments can be identified by the how well they measure competency-constructs, and if those constructs align with what certification-bodies require. This can be achieved by measuring construct validity using CFA and Generalisability theory reliability analyses. Competency construct validity can be reframed as measures of supervisors' capability of adequately assessing trainees' competency by providing the early collection of internal validity evidence for feedback about assessor accuracy in a timeframe and sample size considered minimally sufficient for CFA and to be used for CQI early in a training program year cycle. The use of validity evidence to compare the ability of groups to measure competency constructs was explored by evaluating measurement invariance for binary group comparison. Measurement invariance was analysed to determine if the same variable is being measured across different groups. Reliability was measured using Generalisability Theory with a one facet model for a nested unbalanced data-set to measure score-variance for trainees and the number of assessments for a minimum adequate reliability (NAMAR) value ≥0.80 for each competency-indicator. A minimum number of assessments for CFA and short time-frame were used to explore the possible use in CQI. CFA established that a respecified hypothesised three-competency-domain structure and not the original four-competency-domain structure or initial hypothesised structure provided the best model fit (χ2,p-value,df=29.298,0.2091,24; χ2/df=1.22; TLI=0.994; CFI=0.996; RMSEA=0.032(95%CI 0.000 to 0.067); SRMR=0.019). Competency-item reliability was very poor, all having <20% variance due to trainees, and NAMAR values from 13 to >100. Some were unable to be measured at all. Dimensional equal-form invariance and metric invariance was demonstrated within gender subgroups; equivocal scalar invariance was demonstrated. All were measurable with the sample size in a time frame suitable for CQI for supervisors. Assessment methods proposed by certification-bodies were again not validated. Routine measurement of assessment results' validity can be achieved in a time frame acceptable for CQI and group comparison.

Chapter 6

Variation in judgement-based rating will impact on both learning and decision making for high-stake assessments. Sources of variation need to be identifiable, monitored, fed back, and benchmarked to maintain the quality of the assessors' ratings. However, whether supervisors are capable of assessing the expected competencies is rarely if ever measured. This study explores applying validity evidence using confirmatory factor analysis (CFA) and reliability using the number of assessments for a minimum adequate reliability of *R*=0.80 (NAMAR), combined with statistical process control (SPC) to examine bias for individual supervisor assessments (stringent/lenient assessors), and if present to measure the effect on validity and reliability. The impact of bias was measured by comparing construct validity of nested supervisor groups using measurement invariance; reliability between the whole group and the groups with bias supervisors removed. The groups

were those with and without a tendency to leniency/stringency. SPC application identified outlier assessments (n = 81) and non-outlier assessments (n = 134) as the groups for analysis. Equal form measurement invariance comparing nested model fit of the assessments of outlier supervisors with non-outliers showed an acceptable model fit indicating they are measuring the same competency-domain constructs ($\chi2=66.351$, $p=0.0163$; $\chi2/df=1.51$; df=44; TLI=0.973; CFI=0.983; RMSEA=0.069 (95%CI=0.030-0.101); SRMR=0.083). Equal factor loadings and intercepts for the model had acceptable model fit indices without degradation as indicated by a non-significant difference in the χ^2 with the equal form model ($\chi2=80.898$, $p=0.0204$; $\chi2/df=1.42$; df=57; $\chi2_{diff}=14.547$, $p>0.5$ for $\Delta df=13$; TLI=0.973; CFI=0.983; RMSEA=0.069 (95%CI=0.030-0.101); SRMR=0.083). The reliability was very poor for the competency-indicators but significantly improved with the removal of outlier assessments. The median of a summed NAMAR=60 for the whole group and NAMAR=13 with biased assessors removed (Wilcoxon signed rank test p=0.002). Combining SPC with current methods of CFA is sufficiently fine-grained to measure the impact of bias in assessment results. The methods were able to document equivalence in construct validity of the two groups and significant difference in reliability evidence between those groups using a minimum sample size for the methods. The approach is suitable for providing outcome measures for quality improvement interventions.

Chapter 7

This dissertation starkly illustrates the pervasive and persisting presence of major issues affecting the validity and reliability of supervisor assessment results. Results used for high-stakes decisions. Decades of concerns from past work, now further supported by more fine-grained methodology, indicate the need to address the quality of preparation, implementation and conduct of supervisors' assessments to ensure fairness for trainees and for society. The evidence supports the need for a constructive reappraisal of the status quo of supervisor assessments. Despite the persistence of poor quality assessments, each of the adverse observations can be remediated and better outcomes potentially realized by implementing true evaluative examination of the quality of assessment results and processes. Creating change using validity evidence as a quality improvement method combined with traditional improvement processes appears to be a justifiable process worthy of further exploration during implementation. Firstly, potential improvements in the current system are possible. The accreditation system for training programmes can be made effective simply by implementing current standards, making sure they are implemented using monitoring with feedback and benchmarking. A clear need exists for strict application of standards and regular measurement of compliance using the same or similar methods as detailed in this dissertation. Additionally, the required competencies that medical practitioners are expected to possess by credentialing bodies, and

expressed by the competency-domains and competency-domain indicators, are problematic. Their assessment could be enforced as one option which will require the training of supervisors to adequately assess all the indicator competencies. Developing the competency-domains identified in this dissertation, ones that supervisors can potentially assess adequately, is a better option. Secondly new directions for future research into assessment strategies are crucial next steps. Use of supervisor assessments for measuring the quality of assessors' capabilities to assess competency constructs is clearly achievable. For the supervisor population providing the assessment results for this dissertation, the evidence for gaps for their ability to assess many competency areas is clear. This can be measured routinely and provides opportunities to identify learning methods to improve supervisors' abilities and develop alternative ways of assessing many of the expected competencies.

Samenvatting

Hoofdstuk 1

Je zou denken dat een logisch en essentieel vereiste van onderwijs- en toetsprogramma's is dat supervisors voldoende geschikt, bekwaam en gekwalificeerd zijn om de klinische en professionele prestaties van aiossen tegen een vereiste standaard te beoordelen. Sterker nog, een absolute voorwaarde voor het draaien van een veilig en effectief zorgstelsel waarbinnen het opleiden van nieuwe artsen een wezenlijk onderdeel vormt. Van elke beoordelaar die onderwijs verzorgt voor aiossen zou men verwachten dat hij/zij met een hoge mate van precisie en nauwkeurigheid het prestatieniveau t.a.v. de te leren stof/vaardigheden dat aiossen vertonen, meten. Deze verwachting wordt echter niet ondersteund door een groot deel van de literatuur over toetsing, en meer specifiek over toetsen die op oordeelsvorming berust zijn. Dit baart zorgen gezien het bewijs dat er ligt dat een aanzienlijk deel van de gezondheidszorg gebrekkig en onveilig is, een wereldwijd probleem dat we zien in ontwikkelde landen en vaker in landen met een laag en middelhoog inkomen. De gevolgen van fouten, het nalaten om effectieve zorg te gebruiken, overmatig gebruik van ondoelmatige zorg en het niet in acht nemen van mensen hun waarden en middelen voor hun gezondheid zijn niet gering. De daaraan gekoppelde verspilling van apparatuur, voorzieningen, tijd en menselijke geest brengt onaanvaardbare economische kosten met zich mee voor zorgstelsels. Elk van deze goed gedocumenteerde fouten houdt rechtstreeks verband met de klinische en professionele competenties van artsen, alsmede hun medische en wetenschappelijke kennis, of het nu gaat om een medische fout of overmatig gebruik dan wel onderbenutting van beproefde medische kwaliteitspraktijken. Zorgen over patiëntveiligheid, geografische verschillen in patiëntenzorg die geen verband houden met de medische wetenschap en een gebrekkige "klantenservice" voor patiënten hebben zich gebundeld om samen de bekwaamheid van artsen en zorgstelsels in het algemeen in twijfel te trekken. De toereikendheid van de manier waarop aiossen worden opgeleid en getoetst teneinde hen als arts klaar te stomen voor de onmiddellijke en langdurige klinische praktijk moet in twijfel worden getrokken. Hoewel er enig bewijs bestaat dat meer klinische supervisie en toetsing van aiossen in verband kan worden gebracht met betere patiënt- en onderwijsresultaten, is dit bewijs niet omvangrijk. Deze stand van zaken is begrijpelijk als het geboden onderwijs van wisselende kwaliteit is en metingen van de mate waarin aan de vereiste klinische en

professionele eindtermen is voldaan onnauwkeurig en niet precies zijn. Dientengevolge zullen metingen van de kwaliteit van een verbeteringsinterventie alsook de resultaten van elke onderwijsinterventie gedoemd zijn de huidige staat van onzekerheid voort te zetten. Er is een wetenschappelijke benadering nodig om de onderwijs- en toetsproblemen te boven te komen die het gevolg zijn van zaken zoals het gebruik van toetsen zonder ook aiossen in de praktijk te observeren, het niet opnemen van meerdere perspectieven van bijvoorbeeld mede-aiossen en patiënten en het geen gebruik maken van meetmethoden waarmee klinische resultaten voldoende nauwkeurig kunnen worden voorspeld. Het minimaliseren van onbekwaamheid in de klinische praktijk is een belangrijk en eenvoudig universeel doel van klinische toetsmethoden. Een belangrijke vereiste van een wetenschappelijke benadering is dat methodiek en meting reproduceerbaar zijn. Om het opleiden en beoordelen door supervisors op de werkplek op een wetenschappelijk niveau te krijgen, moet de aanwezigheid van de voor de klinische en professionele praktijk vereiste competenties niet alleen precies en nauwkeurig gemeten worden, maar ook meetbaar zijn. De eenvoudige doelen van het werk voor dit proefschrift waren in de eerste plaats om na te gaan of er behoefte bestaat aan een ingrijpende verandering van de *status quo* door de validiteit van supervisorbeoordelingen aan een kritische analyse te onderwerpen, waarbij de resultaten in de context van huidige kennis werden gepositioneerd en gebruik werd gemaakt van verfijnde betrouwbaarheids- en validiteitsmaten. Een tweede doel was om te proberen innovatieve verandering teweeg te brengen met een exploratieve benadering van validiteitsbewijs bedoeld om in combinatie met traditionele kwaliteitszorgprocessen de kwaliteit te verbeteren. Het werk voor het proefschrift dat staat voor academische complexiteit is in de eerste plaats een vereenvoudiging en standaardisering van wat beschouwd wordt als complexe statistische en data-analytische methoden. In de tweede plaats gebruikt het deze methoden als uitkomstmaten, niet alleen als validiteitsbewijs, maar ook als kwaliteitsindicatoren voor toetsen die op oordeelsvorming berust zijn en als kwaliteitsindicatoren voor de mate waarin beoordelaars in staat zijn om tot dergelijke oordelen over andermans bekwaamheid te komen.

Hoofdstuk 2

Conform het eerste doel voert dit hoofdstuk een kritische evaluatie uit van de betrouwbaarheid van supervisorbeoordelingen en vergelijkt het die betrouwbaarheid met een algemeen gebruikte werkplekbeoordeling, namelijk de Korte Praktijk Beoordeling (KPB). De supervisorbeoordeling is summatief en het onderzoek maakt gebruik van empirische gegevens. De KPB daarentegen is bedoeld om formatief en leerzaam te zijn, en de methode die we gebruikten om de betrouwbaarheid te beoordelen waren de resultaten van een ongepubliceerde systematische review. Het ontwerp en de methode die we voor deze empirische betrouwbaarheidsstudie gebruikten was een binnen de beoordeling geneste aios met behulp van generaliseerbaarheidstheorie. De aan de aios toe te schrijven

variatie in scores voor elk competentie-item dat werd beoordeeld werd geschat door de *minimum-norm quadratic unbiased estimator (MINQUE)* en die scorevariantie werd gebruikt voor de schatting van het aantal beoordelingen dat vereist is om tot een goede betrouwbaarheidswaarde van 0,80 te komen. Voor de Korte Praktijk Beoordeling (KPB) verkregen we dezelfde maten door deze uit alle betrouwbaarheidsstudies die op dat moment beschikbaar waren te destilleren. De variantie in scores van aiossen voor elk competentie-item varieerde van 5,4% voor SEH-vaardigheden tot 37,2% voor communicatievaardigheden, met een gemiddelde over alle competentie-items van 19,3%. Voor het competentie-item "globale beoordeling" bedroeg deze 27,6%. Deze variantiecomponenten vertaalden zich naar respectievelijk 70, 7, 17 en 10 beoordelingen die nodig waren voor een betrouwbaarheid van 0,80. Opvallend was dat de metingen iets beter uit de bus kwamen dan bij de KPB het geval was voor vergelijkbare competenties die getoetst en in de literatuur gepubliceerd waren. Supervisorbeoordelingen kunnen betrouwbaarder zijn dan beoordelingen op basis van KPB's als we de desbetreffende aan aiossen toe te schrijven scorevarianties en vereiste steekproefgrootten vergelijken. Beide beoordelingsvormen geven resultaten die een aanzienlijke onbetrouwbaarheidsgraad hebben. De resultaten staven de stelling dat ondanks het wijdverbreide gebruik ervan, beoordelaarsafhankelijke beoordelingen van klinische competentie contextgevoelig zijn en van persoon tot persoon en van instelling tot instelling variëren. Om op de juiste manier met deze onbetrouwbaarheid van beoordelaarsbeoordelingen om te gaan, blijft het van essentieel belang dat de betrouwbaarheid van beoordelaarsafhankelijke toetsen op de werkplek in de lokale context onderzocht wordt. Een ingrijpende verandering van de *status quo* is dat supervisorbeoordelingen gebruikt zouden kunnen worden als onderdeel van het werkplekbeoordelingsprogramma van aiossen op voorwaarde dat er voldoende beoordelingen plaatsvinden. Alle opleidingsprogramma's die van supervisorbeoordelingen gebruik maken moeten deze informatie standaard aan het accreditatieorgaan verstrekken als bewijs dat hun beoordelingsproces op orde is.

Hoofdstuk 3

De studie voor dit hoofdstuk mat de constructvaliditeit van de resultaten van competentiebeoordelingen door supervisors, mat de betrouwbaarheid van de onderscheiden competentiedomeinconstructen, mat de stabiliteit van constructen over subgroepen heen en positioneerde de resultaten in de erkende literatuur. Constructen werden gedestilleerd met behulp van exploratieve factoranalyse (EFA), de betrouwbaarheid werd vastgesteld door middel van generaliseerbaarheidstheorie, de stabiliteit van factoren werd onderzocht door een EFA uit te voeren van subgroepen en de resultaten werden vergeleken met eerdere soortgelijke studies. De eenheid van analyse was elke beoordeling en omvatte alle beschikbare beoordelingen. De betrouwbaarheid van de onderscheiden constructen werd bepaald door een variantiecomponentenanalyse uit te voeren van de

somscores van aiossen per factor en het aantal beoordelingen dat nodig was om bij gebruik van het construct tot een voldoende betrouwbare beoordeling te komen; de betrouwbaarheidseenheid van analyse was de score per factor voor elke beoordeling. De EFA resulteerde in 3 factoren die slechts 68% van de variantie verklaarden, waarbij factor 1 in de buurt kwam van een "algemeen professioneel functioneren"-competentie, factor 2 een "klinische vaardigheden"-competentie en factor 3 een "professioneel en persoonlijk"-competentie. De variantie in de somscores van aiossen voor de competentie-items per factor bedroeg respectievelijk 40,4%, 27,4% en 22,9%. Het aantal beoordelingen dat nodig was om tot een betrouwbaarheidscoëfficiënt van 0,80 te komen was respectievelijk 6, 11 en 13. De factoroplossing bleef stabiel voor subgroepen van vrouwelijke aiossen, Australische aiossen, het centrale ziekenhuis, chirurgen, gespecialiseerde stafleden, gastartsen en de opsplitsing in afzonderlijke jaren. Artsen als supervisors, mannelijke aiossen en mannelijke supervisors hadden allemaal een verschillende itemgroepering binnen 3 factoren die allemaal een andere competentie-itemdimensionaliteit hadden, wat duidde op de aanwezigheid van factorinstabiliteit binnen enkele subgroepen. Supervisors beoordelen als groep een dominant constructdomein dat vergelijkbaar is met een algemeen-professioneel-functioneren-competentie. Het feit dat de factoroplossing niet stabiel bleef over verschillende populaties supervisors en aiossen heen betekent echter dat subpopulaties aiossen mogelijk anders beoordeeld worden dan andere. Een gebrek aan standaardisering van competentiecriteria voor supervisorbeoordelingen doet twijfel rijzen over de validiteit van deze beoordelingsmethode zoals die momenteel gebruikt wordt. Deze resultaten leveren verder bewijs dat er behoefte bestaat aan verbetering en verandering, vooral omdat de resultaten tientallen jaren van vergelijkbare historische bezorgdheid weerspiegelen, aangezien de geschiedenis zich steeds maar lijkt te herhalen zonder verbetering teweeg te brengen.

Hoofdstuk 4

De uit de voorgaande hoofdstukken verkregen informatie bevestigde dat de overeenstemming in de competentieconstructen tussen verschillende beoordelaars en contexten aan variabiliteit onderhevig is. Verschillende supervisors en beoordelaars hebben vaak tegenstrijdige opvattingen van zowel competentieconstructen als andere latente gedragsconstructen, wat beoordelingsresultaten erg onbetrouwbaar maakt voor *high-stakes* beoordelingen. Overeenstemming over de constructen, nauwkeurigheid en precisie met betrekking tot hoe de competentieconstructen die voor de beoordeling van aiossen worden gebruikt geïnterpreteerd moeten worden is van essentieel belang. Om die reden verrichtten we nader onderzoek van het validiteitsbewijs door de interne validiteit en betrouwbaarheid van de competentieconstructen uit de summatieve, eindesemester supervisorbeoordelingen van aiossen na te gaan. Met behulp van confirmatieve factoranalyse (CFA) onderzochten we de interne constructvaliditeit van de beoordelingen. We

toetsten een verondersteld competentieconstructmodel dat we eerder gedistilleerd hadden en vergeleken deze met tegengestelde modellen van potentiële competentieconstructen, waaronder het competentieconstructmodel van de oorspronkelijke beoordeling. Het veronderstelde model met de competentieconstructen “algemeen professioneel functioneren”, “klinische vaardigheden” en “professioneel handelen” bleek goed bij de gegevens te passen (good model fit), beter dan alle andere modellen. Model-fit-indices waren $\chi 2/df = 2,8$; RMSEA = 0,073 (CI 0,057-0,088); CFI = 0,93; TLI = 0,95; SRMR = 0,039; WRMR = 0,93; AIC = 3879; en BIC = 4018). Het optimale model had voldoende meetinvariantie met geneste analyse van belangrijke populatiesubgroepen, wat de aanwezigheid van volledige metrische invariantie ondersteunde. Betrouwbaarheidsschattingen voor het competentieconstruct “algemeen professioneel functioneren” wezen op een middelenefficiënte en betrouwbare beoordeling van een dergelijk construct (6 beoordelingen voor een $R > 0,80$). Itemhomogeniteit was goed (Cronbachs alpha = 0,899). Andere competentieconstructen zijn middelenintensief, daar er ≥11 beoordelingen nodig waren voor een betrouwbare toetsscore. Desalniettemin zijn de interne validiteit en betrouwbaarheid van op oordeelsvorming beruste beoordelingen van klinische competentie aanvaardbaar, mits de competentieconstructen die beoordelaars daadwerkelijk gebruiken op de juiste manier inzichtelijk worden gemaakt. Supervisors creëren echter nog steeds hun eigen individuele beoordelingsproces, onafhankelijk van de invulling die anderen eraan geven en wat regelgevende organen vereisen. Dit werkt een grote mate van onbetrouwbaarheid en *common method bias* in de hand, waardoor de beoordelingsresultaten geen enkele waarde meer hebben.

Hoofdstuk 5

De validiteit van beoordelingsresultaten kan men zien als maatstaf voor de kwaliteit van het desbetreffende beoordelingsproces. Voor op oordeelsvorming beruste toetsen geldt bovendien dat constructvaliditeit een mogelijke maatstaf is voor de mate waarin beoordelaars in staat zijn om valide en betrouwbare feedback te geven. De kwaliteit van *high-stakes* supervisorbeoordelingen kan worden vastgesteld door te onderzoeken hoe goed zij competentieconstructen meten en of die constructen overeenkomen met hetgeen certificeringsinstanties vereisen. Men kan dit doen door de constructvaliditeit te meten met behulp van CFA en op generaliseerbaarheidstheorie gebaseerde betrouwbaarheidsanalyses. De validiteit van competentieconstructen kan geherdefinieerd worden als een maatstaf voor de mate waarin supervisors in staat zijn om aiossen adequaat op hun competenties te beoordelen. Deze validiteit kan namelijk al vroeg in het proces bewijs van interne validiteit verschaffen dat inzicht biedt in de nauwkeurigheid van beoordelaars binnen de tijd en met een steekproefgrootte die als minimaal voldoende wordt geacht voor het verrichten van een CFA. De verkregen inzichten dienen al vroeg in de jaarcyclus van het opleidingsprogramma te worden gebruikt voor continue

kwaliteitsverbeteringsdoeleinden. Het gebruik van validiteitsbewijs voor het vergelijken van groepen wat betreft de mate waarin zij in staat zijn competentieconstructen te meten werd onderzocht door de meetinvariantie te bepalen zodat we een binaire groepsvergelijking konden doen. We analyseerden de meetinvariantie om te kunnen bepalen of in de verschillende groepen steeds dezelfde variabele werd gemeten. De betrouwbaarheid werd gemeten aan de hand van generaliseerbaarheidstheorie met een één-factormodel voor een geneste ongebalanceerde dataset teneinde de scorevariantie voor aiossen te meten, alsmede het aantal beoordelingen dat vereist was om tot een minimaal aanvaardbare betrouwbaarheidswaarde (NAMAR[14]) van ≥ 0,80 per competentie-indicator te komen. We gebruikten een minimumaantal beoordelingen voor de CFA en een korte tijdsduur om het eventuele gebruik ervan voor continue kwaliteitsverbeteringsdoeleinden te onderzoeken. De CFA wees uit dat een herziene veronderstelde factoroplossing met drie competentiedomeinen het meest passende model was, en niet de oorspronkelijke factoroplossing met vier competentiedomeinen of de factoroplossing die we in eerste instantie verondersteld hadden (χ2, p-waarde, df=29,298, 0,2091, 24; χ2/df= 1,22; TLI=0,994; CFI=0, 996; RMSEA=0,032(95%CI 0,000 tot 0,067); SRMR=0,019). De betrouwbaarheid van competentie-items was zeer matig, daar ze allemaal een aan de aios toe te schrijven variantie van <20% en NAMAR-waardes van 13 tot >100 vertoonden. Enkele konden helemaal niet gemeten worden. Binnen de subgroepen naar geslacht werd dimensionale configurale invariantie en metrische invariantie aangetoond; de scalaire invariantie kwam als dubieus uit de bus. Alle waren meetbaar met de steekproefgrootte en binnen een tijdsduur die geschikt was voor continue kwaliteitsverbetering voor supervisors. De door certificeringsinstanties aanbevolen beoordelingsmethoden werden wederom niet gevalideerd. Het is mogelijk om standaardmetingen van de validiteit van beoordelingsresultaten te verrichten binnen een tijdsbestek dat aanvaardbaar is voor continue kwaliteitsverbeterings- en groepsvergelijkingsdoeleinden.

Hoofdstuk 6

Verschillen in op oordeelsvorming beruste beoordelingen zijn van invloed op zowel het leren als op de besluitvorming bij *high-stakes* toetsen. Variantiebronnen moeten identificeerbaar zijn en gemonitord, teruggekoppeld en gebenchmarkt worden om de kwaliteit van beoordelaars' beoordelingen op peil te houden. Of supervisors in staat zijn om de verwachte competenties te beoordelen wordt echter zelden of nooit gemeten. Deze studie onderzoekt de toepassing van validiteitsbewijs met behulp van confirmatieve factoranalyse (CFA) en de betrouwbaarheid aan de hand van het aantal beoordelingen dat vereist is om tot een minimaal aanvaardbare betrouwbaarheidswaarde van *R*=0,80 (NAMAR) te komen, in combinatie met statistische procescontrole (SPC in het Engels), om na te

[14] NAMAR = Number of assessments for a minimum adequate reliability

gaan of er bij individuele supervisorbeoordelingen (strenge/toegeeflijke beoordelaars) bias optreedt en, zo ja, om het effect hiervan op de validiteit en betrouwbaarheid te meten. De invloed van bias werd gemeten door de constructvaliditeit van geneste supervisorgroepen te vergelijken met behulp van meetinvariantie; oftewel een vergelijking van de betrouwbaarheid van de hele groep met die van de groep zonder de supervisors bij wie bias optrad. Zo ontstonden twee groepen: de beoordelaars met en zonder de neiging om soepel/streng te zijn. De SPC-applicatie bracht twee groepen van beoordelingen in kaart waarop de analyse werd gebaseerd: de uitschieters (n = 81) en de niet-uitschieters (n = 134). Configurale meetinvariantie, waarbij de fit van het geneste model van de beoordelingen van uitschieter-beoordelaars werd vergeleken met die van de niet-uitschieters, liet een aanvaardbare model-fit zien, wat erop duidde dat ze dezelfde competentiedomeinconstructen meten ($\chi2$=66,351, p=0,0163; $\chi2/df$=1.51; df=44; TLI=0,973; CFI=0,983; RMSEA=0,069 (95%CI=0,030-0,101); SRMR=0,083). De gelijke factorladingen en intercepten voor het model vertoonden aanvaardbare model-fit-indices zonder degradatie, zoals bleek uit het niet-significante verschil in de χ^2 bij het configurale model ($\chi2$=80,898, p=0,0204; $\chi2/df$=1,42; df=57; $\chi2_{diff}$=14,547, p>0,5 voor Δdf=13; TLI=0,973; CFI=0,983; RMSEA=0,069 (95%CI=0,030-0,101); SRMR=0,083). De betrouwbaarheid ten aanzien van de competentie-indicatoren was zeer matig, maar deze verbeterde aanzienlijk na verwijdering van de uitschieter-beoordelingen. De mediaan van een gesommeerde NAMAR bedroeg 60 voor de gehele groep en 13 voor de groep zonder beoordelaars bij wie bias optrad (de rangtekentoets van Wilcoxon p=0,002). Het combineren van SPC met huidige CFA-methoden zorgt voor voldoende verfijning om de invloed van bias op beoordelingsresultaten te kunnen meten. Met deze methoden kon op basis van een voor deze methoden minimaal vereiste steekproefgrootte in kaart worden gebracht in hoeverre de twee groepen dezelfde constructvaliditeit hadden en konden significante verschillen in betrouwbaarheidsbewijs tussen die groepen worden vastgesteld. De aanpak is geschikt voor het verschaffen van uitkomstmaten voor op kwaliteitsverbetering gerichte interventies.

Hoofdstuk 7

Dit proefschrift laat goed zien dat er sprake is van grote, wijdverbreide en hardnekkige problemen die van invloed zijn op de validiteit en betrouwbaarheid van de resultaten van supervisorbeoordelingen. Resultaten die gebruikt worden voor *high-stakes* besluiten. Tientallen jaren van bezorgdheid van eerder werk, wat nu verder ondersteund wordt door meer verfijnde methodieken, geven aan dat er meer aandacht nodig is voor de kwaliteit van de voorbereiding, uitvoering en organisatie van supervisorbeoordelingen, zodat rechtvaardigheid voor zowel aiossen als de samenleving kan worden gegarandeerd. Het bewijs staaft de gedachte dat het nodig is de *status quo* omtrent supervisorbeoordelingen op constructieve wijze opnieuw te beoordelen. Hoewel de slechte kwaliteit van

beoordelingen een aanhoudend probleem vormt, kan elk van de negatieve constateringen ongedaan worden gemaakt en kunnen er mogelijk betere resultaten worden geboekt als we de kwaliteit van beoordelingsresultaten en –processen aan een heus evaluerend onderzoek onderwerpen. Het teweegbrengen van verandering door validiteitsbewijs als een kwaliteitsverbeteringsmethode aan te wenden in combinatie met traditionele verbeterprocessen blijkt een verantwoord proces dat nader onderzoek verdient tijdens de implementatie. In de eerste plaats zijn verbeteringen in de huidige systematiek mogelijk. De effectiviteit van het accreditatiesysteem voor opleidingsprogramma's kan heel eenvoudig verbeterd worden door ervoor te zorgen dat er bij de implementatie van huidige standaarden gebruik wordt gemaakt van monitoring met feedback en benchmarking. Er bestaat een duidelijke behoefte aan strikte toepassing van standaarden en regelmatige meting van de naleving hiervan met behulp van dezelfde methoden (of vergelijkbare) als die in dit proefschrift nauwkeurig werden beschreven. Voorts zijn de vereiste competenties waarover artsen door certificeringsinstanties geacht worden te beschikken en die in de competentiedomeinen en competentiedomein-indicatoren tot uiting komen problematisch. Eén mogelijkheid is om de beoordeling van deze competenties af te dwingen, waarvoor het noodzakelijk is dat supervisors getraind worden om alle competentie-indicatoren adequaat te beoordelen. Een betere optie is echter om de in dit proefschrift onderscheiden competentiedomeinen te ontwikkelen, d.w.z. competentiedomeinen waarvan de kans groot is dat supervisors die op de juiste wijze kunnen beoordelen. Ten tweede is het opstellen van nieuwe richtlijnen voor toekomstig onderzoek naar beoordelingsstrategieën een cruciale vervolgstap. Het is duidelijk uitvoerbaar om supervisorbeoordelingen te gebruiken voor het meten van de kwaliteit van de mate waarin beoordelaars in staat zijn om competentieconstructen te beoordelen. Wat de supervisorpopulatie die de beoordelingsresultaten voor dit proefschrift heeft verschaft betreft, is er duidelijk bewijs dat de manier waarop zij vele competentiegebieden beoordelen ondermaats is. Dit kan standaard gemeten worden en biedt kansen voor het ontdekken van leermethoden waarmee supervisors' vaardigheden verbeterd kunnen worden en voor het ontwikkelen van andere manieren om veel van de verwachte competenties te beoordelen.