Hidden in plain sight

Summary

Subjectivity in assessment is gaining increasing respect in the medical education community. The overall goal of this proposed research program is to view subjectivity through the lens of language – what we say and how we say it can provide a window through which we can start to see how clinical supervisors construct opinions and judgments about their learners. Analyzing the language assessors use can deepen our understanding of how they conceptualize competence and performance. Learning how others interpret that language can provide necessary evidence to support the validity of using narrative comments in a way that is credible and defensible.

The **Introduction chapter** sets the stage for the five studies that follow by critically reviewing the literature on the use of written comments in the assessment of learners in health professions education. These written comments can serve as a lens with which to better understand how clinical supervisors subjectively conceptualize residents' performance. Several approaches to analyzing assessment language are described, along with a consideration of what these approaches might offer to our understanding. The overall goals for this thesis were twofold: to determine if narrative comments could be used for assessment in a way that is reliable, credible and has validity for its intended purpose; and to gain a deeper and more nuanced understanding of the language attendings use when assessing their residents in Internal Medicine. Given the educational potential of written assessment comments it is important to explore how comments are constructed, why attendings write the way they do, what their language means to others and what that might say about assessment as a whole. Together, the studies in this thesis form a multiphase, mixed-methods program of research.

The study reported in **Chapter 2** used a database of written comments on Internal Medicine residents' in-training evaluation reports (ITERs) from one program in Canada. We found that it was possible for faculty participants to discriminate between these residents based only on the comments, with excellent inter-rater reliability. Both written comments and numeric scores on the ITER in the first postgraduate year (PGY1) were predictive of performance in PGY3 and comments had fairly high correlations with assigned scores within each year. However, because the ITER scores were already fairly predictive of PGY3 performance the comments did not add additional value when both comments and scores were included in a regression model. Still, this study did show that resident performance can be captured through narrative comments in a reliable way.

To understand *how* faculty were able to rank-order comments so reliably – despite the sometimes vague nature of the language they contain – we conducted the study reported in **Chapter 3**, which involved a constructivist grounded theory analysis of interviews with 24 faculty participants. We used constructivist grounded theory in this study because our goal was to develop a framework to explain the process of interpretation. Faculty did not interpret language at face value; rather, they read between the lines to decode the language in an active process to construct meaning from the comments. Their ability to interpret the comments so reliably suggests a shared under-

standing of a "hidden code" in the language used to describe resident performance. This understanding was not perfect, however, and faculty did express difficulties in interpreting vague, generic and dispositional language. This study raised several critical questions which we then attempted to address in subsequent studies, as follows: (i) To what extent is the code universal versus locally specific, for example to a particular residency program or institution? (Chapter 4); (ii) How much written commentary is required to give a stable impression of a resident using this decoding mechanism? (Chapter 4); (iii) Why does such a "hidden code" exist and what purpose might it be serving? (Chapter 5); (iv) Do residents also know the code and can they interpret the comments effectively and reliably? (Chapter 6).

In **Chapter 4** we addressed the first two questions raised at the end of Chapter 3. To do this we enrolled 24 "outsiders", that is, faculty participants from academic departments of medicine across Canada external to our institution. This time we used as our dataset two cohorts of PGY1 residents' comments; faculty were asked to rank-order a set of residents based on either an entire years' worth of comments or based on only the first three months. We found that it did not matter if faculty were insiders or outsiders – reliabilities remained high, suggesting a degree of universality to the "hidden code" in the comments. Further, using only the first three months of comments yielded comparable reliabilities to using the entire years' worth. Decision studies suggest that acceptable reliability can be achieved using two faculty raters and only the first three months of assessment comments.

In **Chapter 5** we explored the question of *why* attendings write the way they do, with a preponderance of vague, generic and dispositional comments. For this purpose we turned to linguistic pragmatics, which focuses on how people use and understand non-literal language. In particular, we used politeness theory, which posits that people use strategies in their communication in order to allow others (or themselves) to "save face". The pervasive use of "hedging" language suggested to us that writing ITER comments is a face-threatening act, for a number of potential reasons outlined in the chapter. Linguists view politeness and hedging as crucial to enabling smooth social interactions and do not consider these strategies in themselves to be problematic. This may explain, in part, why efforts to prompt faculty to write more balanced and critical comments have been met with little success – this social lubrication is necessary to allow faculty attendings to fulfill their various roles: as teachers, mentors, colleagues and assessors. Faculty's use of politeness strategies may also reflect the culture in which we operate and conformity to the norms that exist in the education context. Yet, although politeness in itself is not necessarily problematic, non-literal language use may lead to misinterpretation – and do not yet know how residents interpret their assessment comments.

This last question, also raised in Chapter 3, is addressed in the final study, reported in **Chapter 6**, which completes the story by exploring residents' understanding of ITER comments. For this purpose we recruited 12 PGY2s from our own IM program and replicated the protocol described in Chapter 2. We found that these residents were able to discriminate between PGY1s based on comments alone with extremely high

inter-rater reliability and that the correlation with faculty rankings of the same residents was nearly perfect. Similar to faculty, residents did not take language at face value but made interpretations and inferences, in much the same way that faculty did. They did not seem to misinterpret the messages despite the pervasive use of non-literal language, which should be reassuring to educators. One unexpected finding was that while residents acknowledged and commented on variability between attendings, they seemed to treat it rather nonchalantly, which may lend some support to the continuing push to embrace more subjectivity in assessment.

The **Discussion chapter** comprises the integration stage of the multiphase, mixed-methods program of research presented in Chapters 2-6. During the process of integration and synthesis the findings from all five studies were considered together and four major themes were identified. First, written assessment comments are useful and should no longer be neglected as valuable sources of data. A re-appraisal of existing research in light of our findings helps explain apparent discrepancies between other researchers' findings and our own. In particular, we problematize the concepts of language specificity and feedback in the setting of assessment. Second, assessment language can be vague but is still decodable – the "hidden code" is accessible. The advantages of studying language in context are also discussed. Third, there is a powerful need to "save face" in assessments, which should not be underestimated. Understanding the social value of politeness can lead to new faculty development approaches. Finally, our findings can lend support to the ongoing push towards embracing subjectivity and collectivity in assessment. This chapter also considers the strengths and limitations of using a mixed-methods approach within a program of research. Finally, implications for practice and future directions for research are presented, including a consideration of the validity evidence supporting the use of written comments in resident assessment.

Samenvatting

In de medische onderwijsgemeenschap wordt steeds vaker aandacht besteed aan subjectiviteit bij toetsing. In brede zin beoogt het aan u voorgelegde onderzoeksprogramma subjectiviteit vanuit een taaloogpunt te benaderen; wat we zeggen en de manier waarop we dat doen kan ons aanknopingspunten aanreiken voor een beter begrip van de manier waarop klinisch begeleiders meningen en oordelen vormen over hun studenten. Door het taalgebruik van beoordelaars te analyseren, wordt mogelijk duidelijker hoe zij zich een beeld vormen van bekwaamheid en functioneren. Meer inzicht in hoe anderen die taal interpreteren kan ons het bewijs leveren dat nodig is om de validiteit ten aanzien van het gebruik van narratieve feedback zodanig te onderschragen dat het geloofwaardig en verdedigbaar is.

Het **inleidend hoofdstuk** bereidt de lezer voor op de vijf daaropvolgende studies middels een kritische uiteenzetting van de literatuur over het gebruik van geschreven feedback bij het beoordelen van studenten in het gezondheidszorgonderwijs. Deze geschreven feedback kan als springplank fungeren naar een beter begrip van de manier waarop begeleiders zich op subjectieve wijze een beeld vormen van het functioneren van artsen in opleiding (aiossen). Er worden verschillende benaderingen beschreven voor het analyseren van "toetstaal", welke vervolgens naar waarde worden geschat in een beschouwing van hun eventuele bijdrage aan ons begrip. Het algemene doel van dit proefschrift was tweeledig: 1) te bepalen of narratieve feedback op een betrouwbare, geloofwaardige en valide manier kan worden aangewend voor toetsdoeleinden, en 2) een beter en genuanceerder begrip te krijgen van het taalgebruik van toeziend artsen bij het beoordelen van artsen in opleiding tot internist. Omdat geschreven feedback op het functioneren een positieve rol kan spelen in het leerproces is het belangrijk te onderzoeken hoe die feedback wordt opgebouwd, waarom toeziend artsen schrijven zoals ze schrijven, wat hun taal betekent voor anderen en welke gevolgtrekkingen ten aanzien van toetsing we daaruit zouden kunnen maken. De studies in dit proefschrift vormen samen een uit diverse fasen bestaand (*multiphase)* onderzoeksprogramma waarbij zowel kwalitatieve als kwantitatieve methoden zijn toegepast (*mixed methods*).

In de in **Hoofdstuk 2** gerapporteerde studie werd gebruik gemaakt van geschreven feedback op de evaluatieverslagen van artsen in opleiding tot internist (ITERs*) van één opleiding in Canada. Onze bevinding was dat deelnemende stafleden in staat waren onderscheid te maken tussen deze aiossen op basis van feedback alleen en daarbij een uitstekende interbeoordelaarsbetrouwbaarheid vertoonden. Zowel de geschreven feedback op de ITERs in het eerste jaar van de vervolgopleiding als de daaraan toegekende numerieke scores bleken voorspellers te zijn van het functioneren in het 3e studiejaar. Ook bleek dat binnen elk jaar de feedback vrij sterk correleerde met de toegekende scores. Omdat de ITER-scores echter al een redelijke voorspellende waarde hadden voor het functioneren in het 3e studiejaar, bleek nadat we feedback en

scores hadden opgenomen in een regressiemodel dat de feedback geen extra waarde toevoegde. Desalniettemin werd met deze studie aangetoond dat het functioneren van aiossen op betrouwbare wijze in beeld kan worden gebracht met behulp van narratieve feedback.

Om erachter te komen hoe het kón dat stafleden de feedback op zo een betrouwbare wijze wisten te rangschikken, ondanks de vage taal die deze soms bevatte, verrichtten we de in **Hoofdstuk 3** vermelde studie waarbij we aan de hand van een constructivistische gefundeerde theoriebenadering interviews met 24 deelnemende stafleden analyseerden. We kozen voor een constructivistische gefundeerde theoriebenadering, omdat we ten doel hadden een kader te ontwikkelen waarmee het interpretatieproces kon worden verklaard. Stafleden bleken de taal niet letterlijk te interpreteren; in plaats daarvan lazen ze tussen de regels door om de taal middels een actief proces te kunnen ontcijferen en betekenis te kunnen abstraheren uit de feedback.

Het feit dat zij zo goed in staat waren de feedback op betrouwbare wijze te interpreteren maakt het aannemelijk dat zij allen een "verborgen taal" beheersten waarmee zij specifiek het functioneren van aiossen beschreven. Deze taalbeheersing was echter niet perfect, daar stafleden te kennen gaven hier en daar moeite te hebben met het interpreteren van vaag, algemeen en persoonsgebonden taalgebruik. De studie riep enkele belangrijke vragen op die we in de volgende studies trachtten te beantwoorden, namelijk: (i) In hoeverre is deze verborgen taal universeel dan wel plaatsgebonden, bijvoorbeeld aan een bepaalde vervolgopleiding of instelling? (Hoofdstuk 4); (ii) Hoeveel geschreven feedback is er nodig om een stabiel beeld te geven van een aios middels dit ontcijfermechanisme? (Hoofdstuk 4); (iii) Waarom bestaat er zo een "verborgen taal" en wat zou het doel ervan kunnen zijn? (Hoofdstuk 5); (iv) Kennen aiossen deze taal ook en weten zij de feedback op efficiënte en betrouwbare wijze te interpreteren? (Hoofdstuk 6).

In **Hoofdstuk 4** werden de eerste twee vragen beantwoord die aan het eind van Hoofdstuk 3 werden gesteld. Daartoe nodigden we 24 "buitenstaanders" uit, dus deelnemende stafleden van academische geneeskundeafdelingen uit heel Canada die niet aan onze instelling verbonden waren. Onze dataset bestond ditmaal uit de feedback van twee cohorten eerstejaars aiossen; we vroegen stafleden een aantal aiossen te rangschikken op basis van alle gedurende een jaar ontvangen feedback of op basis van feedback over alleen de eerste drie maanden. Onze bevinding was dat het niet uitmaakte of stafleden nu binnen- of buitenstaanders waren: de betrouwbaarheid bleef hoog, wat erop duidde dat de "verborgen taal" in de feedback in zekere mate universeel was. Verder maakte het voor de betrouwbaarheid vrijwel geen verschil of de beoordeling berustte op feedback over alleen de eerste drie maanden of op feedback over het gehele jaar. Studies over besluitvorming tonen aan dat een acceptabele be-

trouwbaarheid bereikt kan worden wanneer er twee stafleden worden ingezet als beoordelaar en wanneer de feedback op het functioneren alleen de eerste drie maanden betreft.

In **Hoofdstuk 5** onderzochten we de vraag waarom toeziend artsen schrijven zoals ze schrijven met overwegend vaag, algemeen en persoonsgebonden feedback. Wij zochten het antwoord in de linguïstische pragmatiek die zich richt op hoe mensen niet-letterlijke taal gebruiken en begrijpen. We maakten voornamelijk gebruik van beleefdheidstheorie waarbij gesteld wordt dat mensen strategisch communiceren om te voorkomen dat anderen (of zijzelf) "gezichtsverlies" lijden. Het veelvuldige gebruik van "indekkende" taal deed ons veronderstellen dat het schrijven van ITER-feedback een gezichtsbedreigende aangelegenheid is, om een aantal mogelijke redenen die in het hoofdstuk worden uiteengezet. Taalkundigen beschouwen beleefdheid en indekken als essentieel voor een soepel verloop van sociaal contact en zien geen kwaad in het gebruik van communicatiestrategieën op zichzelf. Dit zou (deels) kunnen verklaren waarom pogingen om stafleden ertoe aan te zetten meer gebalanceerde en kritische feedback te schrijven tot nu toe weinig succes hebben gehad: toeziend artsen hebben deze sociale smering nodig om hun verschillende rollen als docent, mentor, collega en beoordelaar te kunnen vervullen. Het gebruik van beleefdheidsstrategieën door stafleden kan ook een weerspiegeling zijn van de cultuur waarin we ons begeven en van naleving van de normen die in de onderwijscontext gelden. Toch moeten we er rekening mee houden dat, ook al hoeft beleefdheid op zich niet problematisch te zijn, niet-letterlijk taalgebruik verkeerde interpretaties in de hand kan werken. Daarnaast weten we nog niet hoe aiossen zelf de feedback op hun functioneren interpreteren.

Deze laatste vraag die ook in Hoofdstuk 3 werd gesteld, wordt beantwoord in de laatste, in **Hoofdstuk 6** beschreven studie, welke het verhaal afsluit met een nader onderzoek van het begrip dat aiossen hebben van ITER- feedback. Voor dit doel nodigden we 12 tweedejaars aiossen van onze eigen interne geneeskundeopleiding uit en voerden we dezelfde stappen uit als die we in Hoofdstuk 2 beschreven. Onze bevinding was dat deze aiossen in staat waren onderscheid te maken tussen eerstejaars aiossen op basis van alleen feedback en daarbij een buitengewoon hoge interbeoordelaarsbetrouwbaarheid vertoonden. Bovendien was de correlatie met de rangorde die stafleden gemaakt hadden van dezelfde aiossen vrijwel perfect. Net als de stafleden namen de aiossen de feedback niet letterlijk, maar interpreteerden deze en trokken conclusies op nagenoeg dezelfde wijze als stafleden dat deden. Niets wees erop dat zij het commentaar verkeerd interpreteerden, ondanks het veelvuldige gebruik van niet-letterlijke taal, en dit zou opleiders gerust moeten stellen. Wat we niet verwachtten te vinden was dat aiossen nogal onverschillig stonden ten opzichte van de verschillen tussen toeziend artsen, terwijl zij wel degelijk erkenden dat ze bestonden en dit ook opmerk-

ten. Deze bevinding kan de aanhoudende vraag naar meer subjectiviteit bij toetsing kracht bijzetten.

Het **Discussiehoofdstuk** beslaat de laatste fase van het *multiphase*, *mixed-methods* onderzoeksprogramma tijdens welke de in Hoofdstukken 2 t/m 6 gepresenteerde fasen worden geïntegreerd. Tijdens het integratie- en syntheseproces werden de bevindingen van alle vijf de studies in hun geheel beschouwd en daarbij kwamen vier hoofdthema's naar voren. Ten eerste is geschreven feedback op het functioneren bruikbaar en deze zou als waardevolle gegevensbron niet meer onbeschouwd mogen blijven. Door bestaand onderzoek opnieuw te bekijken in het licht van onze bevindingen, kunnen duidelijke afwijkingen tussen de bevindingen van andere onderzoekers en de onze worden verklaard. Zo maken we bijvoorbeeld teveel een probleem van concepten als taalspecificiteit en feedback als het gaat om toetsing. Ten tweede mag toetstaal dan wel vaag zijn, maar het is toch te ontcijferen; met andere woorden: men heeft toegang tot de "verborgen taal". De voordelen van het bestuderen van taal in de context worden ook besproken. In de derde plaats bestaat er een sterke behoefte aan het voorkomen van "gezichtsverlies" bij toetsing, welke niet onderschat moet worden. Een beter begrip van de sociale waarde van beleefdheid kan ons helpen docentprofessionaliseringsprogramma's te vernieuwen. Ten slotte kunnen onze bevindingen het draagvlak voor subjectiviteit en collectiviteit bij toetsing verbreden. Dit hoofdstuk bespreekt ook de voor- en nadelen van een *mixed methods*-benadering binnen een onderzoeksprogramma. Als laatste worden de gevolgen voor de praktijk en aanbevelingen voor toekomstig onderzoek gepresenteerd, samen met een beschouwing van het bewijs dat de validiteit van toetsing van aiossen op basis van geschreven feedback bij de beoordeling van aiossen aantoont en het gebruik daarvan ondersteunt.