

## Arbeitsplatz-basiertes Assessment im Medizinstudium: Ist eine Multilevel-Analyse sinnvoll?

Wolfgang Himmel  
Institut für Allgemeinmedizin der  
Universitätsmedizin Göttingen  
Humboldtallee 38  
Göttingen  
whimmel@gwdg.de

Armin Gemperli  
Kultur- und Sozialwissenschaftliche  
Fakultät Luzern  
Frohburgstraße 3  
Luzern  
armin.gemperli@unilu.ch

Anja Rogausch  
Institut für medizinische Lehre  
Medizinische Fakultät Universität Bern  
Konsumstraße 13  
Bern  
anja.rogausch@iml.unibe.ch

### Zusammenfassung

Hintergrund. Einschätzungen der praktisch-klinischen Fähigkeiten von Studierenden (z. B. Anamnese- u. Gesprächsfähigkeiten) finden an vielen Universitäten mittlerweile regelmäßig statt: zur Messung des individuellen Lernerfolgs, aber auch zur Feststellung von Unterschieden in der didaktischen Qualität von Lehrenden. Studierende werden durch eine oder mehrere Personen in mehreren Kliniken/Instituten bzw. medizinischen Fächern zumeist mehrfach bewertet. Einfache Auswertungen dieses Materials können zu Fehlurteilen führen, da die Einschätzungen nicht unabhängig voneinander sind. Das ist als generelles Phänomen unter dem Begriff Mehrebenen- oder Multilevel-Struktur bekannt, wird aber eher selten berücksichtigt – u. a. weil entspr. Statistik-Programme für dieses Problem nicht ganz einfach in der Anwendung sind.

Ziel. Anhand eines umfangreichen Datensatzes von Mehrfach-Bewertungen von 165 Medizinstudierenden in verschiedenen Kliniken soll gezeigt werden, wie mit dem Programm *proc mixed* von SAS der Multilevel-Charakter der Daten berücksichtigt werden kann.

Spezifische Fragestellung und Methode. Es sollte untersucht werden, wovon die Fremdeinschätzungen von Medizinstudierenden im Rahmen sog. Mini Clinical Evaluation Exercises (Mini-CEX) mitbeeinflusst sind. Dafür wurden in der Prozedur *proc mixed* sog. *Fixed effects* und *random effects* eingeführt („gemischtes Modell“), um ihre Wirkung auf die Ergebnisse zu zeigen.

Ergebnisse. Im gemischten Modell waren mehrere Variablen signifikante Prädiktoren für die Gesamteinschätzung der studentischen Leistungen. Zum Beispiel gaben Assistenzärzte deutlich bessere Einschätzungen für die Studierenden – im Vergleich zu Ober- und Chefärzten. Zurückhaltender in den Einschätzungen, also kritischer gegenüber den Studierenden waren die Ergebnisse aus kleineren Kliniken und bei geringerer Komplexität der zu bewertenden Sachverhalte. Im linearen Modell wären dagegen weitere Faktoren – fälschlicherweise – als signifikant markiert worden.

Diskussion. Die Auswertung über *proc mixed* hatte den Vorteil, dass wir uns vor „Signifikanzfällen“ schützen konnten. Nicht ganz einfach war das Verständnis und die Definition der *fixed effects* im Vergleich zu den *random effects*. Hier empfanden wir die Beratung durch den statistischen Experten als besonders wichtig und hilfreich.

**Schlüsselwörter:** Multilevel-Analyse, SAS-Prozedur, PROC MIXED, Fixed und Random Effects, Assessment, Medizinische Ausbildung

## 1 Einleitung

Für viele SAS-Nutzer bzw. Anwender von Statistikprogrammen scheint es eine gewissermaßen unsichtbare Grenze zu geben, was die Nutzung komplexerer Programme bzw. statistischer Verfahren betrifft. So werden zum Beispiel multiple logistische Regressionen auch von Nicht-Statistikern noch häufig gerechnet; anders aber sieht es bei sog. Multilevel- oder Mehrebenen-Analysen aus. Deren Wert bzw. Notwendigkeit wird oft betont; selten dagegen trauen sich durchschnittliche Nutzer von Statistikprogrammen an die Durchführung entsprechender Auswertungen. Diese „Zurückhaltung“ scheint uns durchaus ein sinnvolles Thema auf einer KSFE.

Ziel des folgenden Beitrages ist es, exemplarisch eine entsprechende Auswertung durchzuführen. Dabei verzichten wir bewusst auf statistische „Finessen“ oder Komplikationen. Vielmehr geht es uns darum, einen praktikablen und gut nachvollziehbaren Ansatz für eine typische Multilevel-Datenstruktur zu zeigen. Allerdings besteht die Autorengruppe „zur Sicherheit“ zumindest aus einem Statistiker, wenn auch mit diesem Beitrag der Anspruch verbunden ist, ähnliche Analysen letztlich ohne eine solche Hilfestellung durchzuführen.

## 2 Fragestellung

Die inhaltliche Fragestellung, die den Anlass zu dieser Studie gegeben hatte, spielt für die Tagung eine sekundäre Rolle. Sie sei deshalb nur kurz erwähnt. Wir waren erstaunt über die extrem gute Gesamtbewertung von Studierenden und zukünftigen Ärzten einer schweizerischen Universität – mit einem Durchschnitt von 9,0 (SD 1,0) auf einer 10-Punkte-Skala. Wir wollten daher Faktoren bzw. Prädiktoren für die Leistungsbewertung finden – unter Berücksichtigung der Multilevel-Struktur der Daten.

Die spezifische Fragestellung im Rahmen dieser Tagung war, wie sich konzeptionell und praktisch die Multilevel-Struktur berücksichtigen lässt – mit drei Unterzielen:

- Erarbeitung eines hierarchischen Modells
- Umsetzung als SAS-Syntax
- Verständnis der Ergebnisse und ihre Interpretation.

## 3 Methode

### 3.1 Beschreibung des Datensatzes

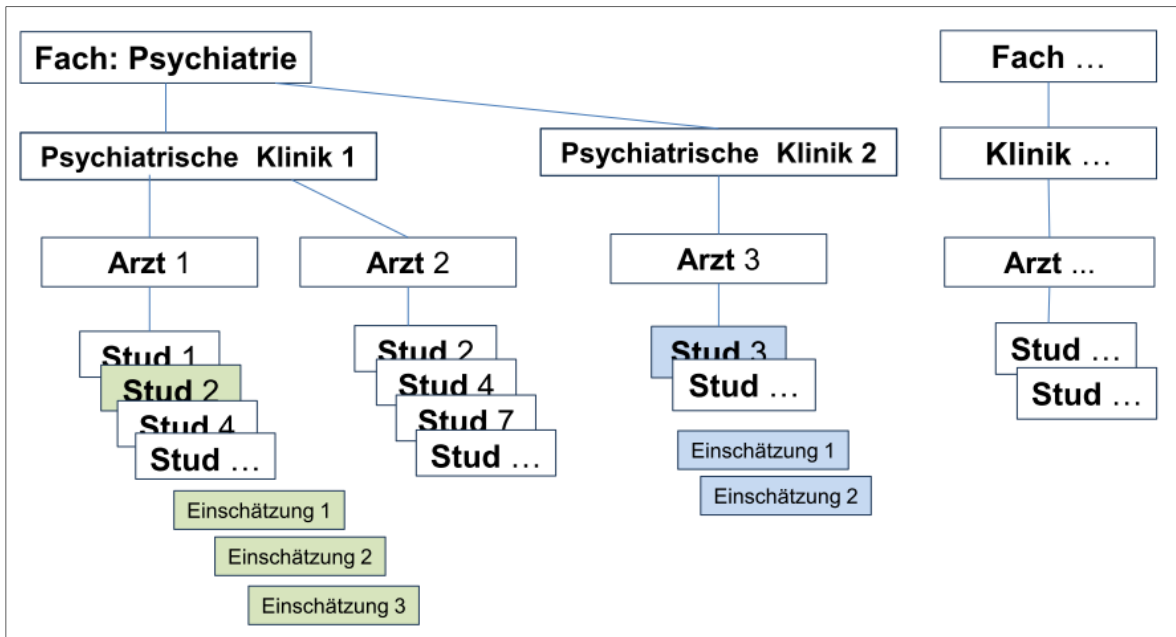
Es lagen uns insgesamt knapp 1.600 Assessments im Rahmen von sog. Arbeitsplatzbasierten Assessments im Medizinstudium vor. Solche Assessments sind mittlerweile in vielen Medizin-Studiengängen Standard. Sie sollen an verschiedenen Stellen der Ausbildung (zum Beispiel in Blockpraktika oder auch später während des praktischen Jahrs bzw. der Assistentenzeit) vor allem die praktischen Fertigkeiten von Studierenden oder Ärzten in Weiterbildung messen. Ein typisches Instrument dafür sind die sog. Mini-Clinical Evaluation Exercises (Mini-CEX). Auf wenigen Beurteilungsdomänen sollen lehrende Ärzte die Fähigkeiten von Studierenden und Ärzten in Weiterbildung bewerten. Insbesondere geht es dabei um die Anamnese, körperliche Untersuchung, Beratung von Patienten und Urteilsfähigkeit bei der Diagnosefindung, aber auch um die Effizienz der Auszubildenden (Norcini & Burch; 2007). Typischerweise schließt sich zum Abschluss noch eine Frage zum Gesamteindruck der Studierenden an, die – wie die anderen Fragen auch – auf einer 10-Punkt-Skala zu beantworten sind. Wir werden uns hier auf die Auswertung der Frage zum Gesamteindruck beschränken.

Die Daten, auf die wir uns im Folgenden beziehen, stammen von der medizinischen Fakultät der Universität Bern. Die Assessments – und damit sind wir nun bei der Besonderheit des Datensatzes – beziehen sich nicht auf 1.600 verschiedene Studierende, sondern zum Teil auf dieselben Personen (insgesamt  $n = 165$ ), stammen aus verschiedenen Kliniken, die wiederum zu bestimmten Fächern gehören und werden teils von denselben, teils von anderen Ärzten durchgeführt. Abbildung 1 zeigt exemplarisch die Multilevel-Struktur.

### 3.2 Modell

Entscheidend für ein hierarchisches bzw. Multilevel-Modell war das richtige Verständnis und die präzise Definition von sog. *fixed effects* und *random effects*. Abbildung 2 gibt eine Definition dieser Effekte, angelehnt an die Literatur (Snijders 2005; Twisk 2013).

Dabei haben wir mit einer sehr einfachen Annahme bzw. Variante gearbeitet, die sich in der schematischen Graphik in Abbildung 2 andeutet: die Annahme nämlich, dass die Studierenden von den zu untersuchenden Einheiten (z.B. Kliniken oder Lehrenden) jeweils spezifisch, also etwas besser oder schlechter, bewertet werden.



**Abbildung 1:** Arbeitsplatz-basiertes Assessment im Medizinstudium: ein Multilevel-Problem

- Fixed effects
  - Alle möglichen Einheiten einer Ebene [z.B. Kliniken] sind einbezogen, „fehlerfrei“ messbar und universell gültig
  - Die Effekte aller Einheiten [z.B. Kliniken] werden geschätzt, z.B. als Dummy-Variable
- Random effects
  - Einheiten sind aus einer Stichprobe gezogen, über die etwas generalisierend ausgesagt werden soll.
  - Problem: jede der Einheiten [Kliniken; Lehrende] hat eine Tendenz, besser oder schlechter zu beurteilen
  - Für die Varianz zwischen den Einheiten wird ein Parameter geschätzt (als „random“)

**Abbildung 2:** Multi-Level-Modell: random effects and fixed effects

Für unseren spezifischen Datensatz haben wir dann die verschiedenen Variablen als *fixed effects* oder *random effects* definiert (Abbildung 3). Hier sei schon angemerkt, dass es keine eindeutige Regel für eine solche Definition gibt. Nicht zuletzt ist die Entscheidung, ob eine Variable ein *fixed* oder *random effect* ist, von der Fragestellung und der Messung abhängig. Würde man, um ein Beispiel zu nennen, in unserem Fall Interesse daran haben, ob und wie sich ein Fach, zum Beispiel die Chirurgie, von allen anderen Fächern generell unterscheidet, hätte man an der betreffenden Universität versuchen können, die Ergebnisse der Chirurgie allen anderen Fächern gegenüberzustellen und das gewissermaßen als *fixed effect* zu formulieren.

- Fixed effects
  - Geschlecht der Studierenden und Ausbilder
  - Schwierigkeit der Assessments [hoch, mittel, gering]
  - Ergebnis einer Vorprüfung im Juni [hoch, mittel, gering]
  - Funktion / Hierarchie der Lehrenden [Assistenten, Oberärzte, Chefärzte]
  - Klinikgröße [groß, mittel, klein]
  
- Random effects
  - Lehrende
  - Studierende
  - Medizinische Spezialität [Fach]
  - Klinik

**Abbildung 3:** Fixed effects und random effects in unserem Datensatz

### 3.3 Umsetzung als SAS-Programm

Üblicherweise wird ein solcher Datensatz mit dieser Fragestellung durch ein sog. „gemischtes Modell“ ausgewertet, was mit der Prozedur *proc mixed* in SAS geschätzt werden kann (Little et al. 2006). Der entsprechende Code befindet sich in Abbildung 4. Die SAS-Syntax für diese Prozedur ist vergleichbar mit der Syntax zur logistischen Regression (*proc logistic*) oder zum allgemeinen linearen Modell (*proc glm*). Alle im Model-Statement nach dem Gleichheitszeichen genannten Variablen sind die *fixed effects*. Am Ende des ersten Absatzes werden für die Ausgabe der Schätzer auch die Konfidenzintervalle angefordert (*solution cl*).

Im zweiten Absatz des Statements werden die *random effects* definiert, nach dem Ausdruck „subject=“. Davor ist der Typus des *random effects* definiert. In diesem Fall, wie oben bereits angedeutet, handelt es sich um die einfachste Variante, nämlich ein „Intercept“-Modell – also die Annahme, dass die untersuchten Einheiten (Lehrende, Kliniken, etc.) die Studierenden jeweils etwas besser oder schlechter bewerten (dargestellt durch die Y-Achse parallel schneidenden Geraden).

Im letzten Absatz des Programms schließlich werden für die *fixed effects* die kleinsten Quadrate-Lösungen (mit Vertrauensintervall) angefordert. Man kann gewissermaßen lebendig sehen, wie sich – unter Beachtung der *random effects* – die Mittelwerte für alle Ausprägungen der einzelnen Variablen (z.B. Klinikgröße, Funktionshierarchie) darstellen.

```

proc mixed data=IML method=ml;
  class  specialty_num n_stud_gruppen3 function student_gender
        student_id expert_id clinic complexity CS_Juni_3g;
  model  e_overall_n = n_stud_gruppen3 function
        student_gender complexity CS_Juni_3g /
        solution CL ;

  random intercept / subject=specialty_num ;
  random intercept / subject=clinic;
  random intercept / subject=expert_id;
  random intercept / subject=student_id;

  lsmeans n_stud_gruppen3 / cl;
  lsmeans function / cl;
  lsmeans student_gender / cl;
  lsmeans complexity / cl;
  lsmeans CS_Juni_3g / cl;

run;

```

Abbildung 4: Umsetzung als SAS-Programm

## 4 Ergebnisse

Abbildung 5 zeigt zwei wesentliche Teile des Outputs bei *proc mixed* (hier nur als Ausschnitte). Im oberen Teil sind die jeweiligen Schätzer für den Intercept und die Abweichung für die einzelnen Ausprägungen bei den Variablen zu sehen, einschließlich (rechts außen) die Vertrauensintervalle. Noch leichter lesbar ist der untere Teil des Outputs, bei dem für alle Variablen-Ausprägungen die jeweiligen Mittelwerte, für die *random effects* kontrolliert, gezeigt werden (Ergebnis der *ls means*-Anweisung).

Lösung für feste Effekte													
Effekt	n_stud_gruppen3	qualification of teacher (1= Assistenzarzt, 2= Oberarzt, 3= Chefarzt)	gender of student (1=Mann, 2=Frau)	Komplexität der Aufgabe (1=gering, 2=mittel, 3=hoch)	voriges Prüfungsergebnis (1=gering, 2=mittel, 3=hoch)	Schätzwert	Standardfehler	DF	t-Wert	Pr >  t	Alpha	Untere	Obere
Intercept						8.9662	0.1753	4	51.15	<.0001	0.05	8.4795	9.4528
n_stud_gruppen3	1-15 Stud					-0.2606	0.1048	528	-2.49	0.0132	0.05	-0.4664	-0.05475
n_stud_gruppen3	16-30 Stud					-0.1706	0.1158	528	-1.47	0.1414	0.05	-0.3982	0.05695
n_stud_gruppen3	über 30					0	.	.	.	.	0.05	.	.
function		Assistenzarzt				0.5494	0.1452	528	3.78	0.0002	0.05	0.2642	0.8347
function		Oberarzt				0.1205	0.1499	528	0.80	0.4219	0.05	-0.1740	0.4150
function		Chefarzt				0	.	.	.	.	0.05	.	.

Kleinste-Quadrate-Mittelwerte													
Effekt	n_stud_gruppen3	qualification of teacher (1= Assistenzarzt, 2= Oberarzt, 3= Chefarzt)	gender of student (1=Mann, 2=Frau)	Komplexität der Aufgabe (1=gering, 2=mittel, 3=hoch)	voriges Prüfungsergebnis (1=gering, 2=mittel, 3=hoch)	Schätzwert	Standardfehler	DF	t-Wert	Pr >  t	Alpha	Untere	Obere
n_stud_gruppen3	1-15 Stud					8.7011	0.08666	528	100.41	<.0001	0.05	8.5308	8.8713
n_stud_gruppen3	16-30 Stud					8.7910	0.09804	528	89.67	<.0001	0.05	8.5984	8.9836
n_stud_gruppen3	über 30					8.9616	0.09617	528	93.18	<.0001	0.05	8.7727	9.1506
function		Assistenzarzt				9.1440	0.06924	528	132.07	<.0001	0.05	9.0080	9.2800
function		Oberarzt				8.7151	0.08053	528	108.22	<.0001	0.05	8.5569	8.8733
function		Chefarzt				8.5946	0.1408	528	61.06	<.0001	0.05	8.3181	8.8711

Abbildung 5: Output für *proc mixed* (Ausschnitt)

In Abbildung 6 sind die beiden Informationen aus den Outputs zusammengefasst: Man sieht also den Schätzer und die jeweiligen Abweichungen bei den Variablen-Ausprägungen bzw. im rechten Teil der Abbildung die *random*-kontrollierten Mittelwerte für alle Ausprägungen.

Prädiktor	Kleinste-Quadrate-Mittelwerte			
	Schätzer	(95%-Vertr.-Interv.)	Schätzer	(95%-Vert.-Interv.)
Intercept	8,96			
Chefarzt	0		8,5	(8,3 ; 8,8)
Oberarzt	0,12	(-0,17 ; 0,41)	8,7	(8,5 ; 8,8)
Assistenzarzt	0,54	<b>(0,26 ; 0,83)</b>	9,1	(9,0 ; 9,2)
Klinik (groß)	0		8,9	(8,5 ; 8,8)
Klinik (mittel)	-0,17	(-0,39 ; 0,05)	8,7	(8,5 ; 8,9)
Klinik (klein)	-0,26	<b>(-0,46 ; -0,05)</b>	8,7	(8,7 ; 9,1)
Stud. ♀	0		8,8	(8,7 ; 8,9)
Stud. ♂	0,04	(-0,19 ; 0,07)	8,7	(8,6 ; 8,9)
Komplexität (hoch)	0		8,9	(8,7 ; 9,0)
Komplexität (mittel)	-0,08	(-0,21 ; 0,05)	8,8	(8,7 ; 8,9)
Komplexität (gering)	-0,23	<b>(-0,44 ; -0,19)</b>	8,6	(8,4 ; 8,8)
Vorh. Prüfungsleistung (hoch)	0		8,9	(8,7 ; 9,0)
Vorh. Prüfungsleistung (mittel)	-0,12	(-0,27 ; 0,01)	8,7	(8,7 ; 8,9)
Vorh. Prüfungsleistung (gering)	-0,14	(-0,36 ; 0,06)	8,7	(8,4 ; 8,8)

**Abbildung 6:** Ergebnisse von *proc mixed*

Auch wenn, wie oben angedeutet, die spezifischen Inhalte unserer Studie auf dieser Tagung keine Rolle spielen, sei darauf hingewiesen, dass zumindest einige „Faktoren“ signifikant im Modell waren. Zum Beispiel gaben Assistenzärzte deutlich bessere Einschätzungen für die Studierenden – im Vergleich zu Ober- und Chefarzten. Zurückhaltender in den Einschätzungen, also kritischer gegenüber den Studierenden waren die Ergebnisse aus kleineren Kliniken und bei geringerer Komplexität der zu bewertenden Sachverhalte. Signifikante Ergebnisse sind durch die fettmarkierten Vertrauensintervalle markiert.

Interessant ist in unserem Zusammenhang die Frage, ob sich die Ergebnisse durch das gemischte Modell – im Vergleich zu einem linearen Modell – unterscheiden. In der Tat ist dies der Fall. So würden zum Beispiel im linearen Modell auch die Faktoren „mittlere Kliniken“ und „mittlere Vor-Prüfungsleistungen“ signifikant sein (Daten nicht gezeigt). Übrigens lässt sich dieser Vergleich sehr einfach anstellen, indem man die „random“ Statements im SAS Programm auskommentiert. Dann entspricht *proc mixed* der Prozedur *proc glm*.

Für Nutzer ist natürlich auch interessant, die differenzielle Wirkung bzw. die Effektstärke der *random effects* annähernd abschätzen zu können. Hier empfiehlt sich die Be-

trachtung der im Output dargestellten sog. Kovarianzparameterschätzer (Abbildung 7). So würde in unserem Fall der Faktor „Experte“ (= Lehrende) am stärksten zu Buche schlagen, fast ähnlich stark wie der Fehler („residual“).

Kovarianzparameterschätzer		
Kov.Parm	Subjekt	Schätzwert
Intercept	specialty_num	0
Intercept	clinic	0.01280
Intercept	expert_id	0.3702
Intercept	student_id	0.05411
Residual		0.4690

Abbildung 7: Effektstärke der random effects

## 5 Diskussion

Soweit es um die inhaltliche Fragestellung ging, haben die Funktionshierarchie der Ärzte, die Komplexität der Anforderung und die Größe der Klinik einen gewissen Einfluss auf die Bewertung der studentischen Leistungen. Dass dieser Einfluss vom Effekt her eher gering war, hat natürlich auch damit zu tun, dass das Gesamtniveau der Bewertungen extrem hoch lag (*ceiling*-Effekt). Erstaunlich für uns war, dass die Prüfungsvorleistungen (als vermeintlicher Indikator der Kenntnisse von Studierenden) nur geringfügige Auswirkungen auf die Einschätzungen der Fähigkeiten der Studierenden hatten.

Die inhaltlich gerechtfertigte Multilevel-Auswertung über `proc mixed` zeigte – im Vergleich zur linearen Auswertung durch `proc glm` – kaum sichtbare bzw. keine starken Auswirkungen auf die Ergebnisse. Im ersten Moment könnte man denken, es handele sich also um eine *l'art pour l'art*-Übung. Dem ist nicht so! Hätten wir die gesamten ca. 1.600 Bewertungen von Studierenden als unabhängige Messungen betrachtet, hätten unsere Daten einen viel zu großen Stichprobenumfang vorgespiegelt und wir deren Informationsgehalt überschätzt. Tatsächlich bestand die Stichprobe von 1.588 Datenpunkten aus „nur“ 165 Individuen. Zu diesen liegen Mehrfachmessungen vor, von denen man zunächst einmal annehmen muss, dass es sich um jeweils leichte variierte Messungen handelt. Das betrifft nicht nur Mehrfachmessungen beim selben Subjekt, sondern auch Messungen durch den jeweils selben Arzt (Lehrenden) oder in derselben Kliniken etc. Genau dies wird im Multilevel-Model durch random effects berücksichtigt. Praktisch bedeutet dies, dass wir den unserer Stichprobe innewohnenden Informationsgehalt realistisch benennen. Statistisch bedeutet es, dass die Vertrauensintervalle der



Effektschätzer breiter werden und wir uns vor „Signifikanzfallen“ besser schützen können. Eine Auswertung im Sinne des linearen Modells hätte zumindest zwei weitere Faktoren (Prädiktoren) fälschlicherweise als signifikant erscheinen lassen. Bei etwas anders gelagerten Daten (z. B. ohne bzw. nur geringem ceiling-Effekt) wären noch deutlichere Fehleinschätzungen durch ein falsch gewähltes statistisches Modell zu erwarten gewesen.

Als Limitationen sollten wir zumindest stichwortartig erwähnen, dass wir den Zeit-Charakter der Messungen nicht berücksichtigt haben. Dies wäre in der Prozedur *proc mixed* natürlich möglich. In Anbetracht des Anspruchs einer zunächst unkomplizierten Lösung haben wir darauf verzichtet. Ebenso haben wir auf andere Annahmen über den Charakter der *random effects* verzichtet und sind von einer vergleichsweise einfachen Annahme – linear unterschiedliche Levels der Bewertungseinheiten – ausgegangen.

Mit unserer Auswertung und dem Vortrag haben wir weder neue Auswertungsmethoden zur Diskussion gestellt noch besonders elegante Lösungen für bekannte Probleme dargestellt bzw. weiterentwickelt. Wir haben letztlich eine bekannte und vergleichsweise schlichte Auswertung eines Mehrebenen-Datensatzes vorgeführt. Dies erschien uns auf einer Konferenz der SAS-Nutzer deshalb wichtig, weil vor allem in der Medizin die Anwendungsfälle von Multilevel-Auswertungen in Zukunft sicher noch häufiger werden.

Während Multilevel-Analysen gerade in schulischen Umgebungen (geprägt durch verschiedene Lehrer und Klassen) entwickelt wurden und zum Einsatz kommen, konnten wir exemplarisch zeigen, dass Untersuchungen zur Ausbildung von Medizin-Studierenden eher noch komplexer sind, d. h. mehr Ebenen bedeutsam werden können. Ein weiteres, in der Zukunft sicher zunehmendes Feld für solche Auswertungen werden sogenannte Multicenter-Studien sein. Sie sind natürlich jetzt schon typisch in der Medizin, werden aber gerade z. B. in der hausärztlichen Medizin häufiger werden, weil interessante Ergebnisse für die praktische Versorgung in der Medizin zu erwarten sind. Aber für valide Aussagen über Allgemeinpraxen und die dort behandelten bzw. versorgten Patienten ist die Mehrebenen- bzw. Multilevel-Struktur zu berücksichtigen. Das geschieht bisher noch eher selten (als positives Beispiel sei eine kürzlich veröffentlichte Untersuchung von Schäfer et al. 2012 genannt).

Mit dem hier vorgestellten sehr einfachen Modell hoffen wir gezeigt zu haben, dass die vornehme Zurückhaltung gegenüber der Anwendung gemischter Modelle durch Nicht-Statistiker nur bedingt begründet ist. Zumindest ist für einen unkomplizierten Standardfall die Syntax bei dieser Prozedur nicht anspruchsvoller als bei anderen Prozeduren, die von statistischen Laien oft genutzt werden. Vergleichsweise schwieriger für uns war ein Verständnis und die Definition der *fixed effects* im Vergleich zu den *random effects*. Hier empfanden wir die Beratung durch den statistischen Experten als besonders wichtig und hilfreich. Umgekehrt hat die vergleichsweise einfache Syntax für uns den Vorteil, dass wir „spielerisch“ – seriöser formuliert: explorierend – weitere Faktoren aus

dem Datensatz werden einfügen und in ihrer Wirkung abschätzen können, ohne für solche Datenauswertungen immer Rücksprache mit den statistischen Experten halten zu müssen.

## **Literatur**

- [1] J. Norcini, V. Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teach*; 29: 855-71. 2007
- [2] R. C. Littell, G. A. Milliken, W. W. Stroup et al. *SAS for Mixed Models*, 2<sup>nd</sup> ed. Cary, NC: SAS Institute Inc., 2006.
- [3] I. Schäfer, H. Hansen, G. Schön et al. The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. first results from the multicare cohort study. *BMC Health Serv Res*; 12: 89. 2012.
- [4] T. A. B. Snijders. Fixed and Random Effects. *Encyclopedia of Statistics in Behavioral Science*; vol. 2, 664-665. Chicester: Wiley, 2005.
- [5] J. W. R. Twisk. *Applied Multilevel Analysis: A Practical Guide*, 6<sup>th</sup> ed., Cambridge Univ. Pr., Cambridge 2013.