ORIGINAL ARTICLES

Generalizability of a Composite Student Selection Procedure at a University-Based Chiropractic Program

Lotte D. O'Neill, MMedEd, DC, University of Southern Denmark, **Lars Korsholm**, PhD, University of Southern Denmark, **Birgitta Wallstedt**, MSc, University of Southern Denmark, **Berit Eika**, PhD, MD, MHPE, MI, University of Aarhus, **Jan Hartvigsen**, PhD, DC, University of Southern Denmark, Nordic Institute of Chiropractic and Clinical Biomechanics

Purpose: Non-cognitive admission criteria are typically used in chiropractic student selection to supplement grades. The reliability of non-cognitive student admission criteria in chiropractic education has not previously been examined. In addition, very few studies have examined the overall test generalizability of composites of non-cognitive admission variables in admission to health science programs. The aim of this study was to estimate the generalizability of a composite selection to a chiropractic program, consisting of: application form information, a written motivational essay, a common knowledge test, and an admission interview. **Methods:** Data from 105 Chiropractic applicants from the 2007 admission at the University of Southern Denmark were available for analysis. Each admission parameter was double scored using two random, blinded, and independent raters. Variance components for applicant, rater and residual effects were estimated for a mixed model with the restricted maximum likelihood method. The reliability of obtained applicant ranks (generalizability coefficients) was calculated for the individual admission criteria and for the composite admission procedure. **Results:** Very good generalizability was found for the common knowledge test (G=1.00) and the admission interview (G=0.88). Good generalizability was found for application form information (G=0.75) and moderate generalizability (G=0.50) for the written motivation essay. The generalizability of the final composite admission procedure, which was a weighted composite of all 4 admission variables was good ($G_c = 0.80$). **Conclusion:** Good generalizability for a composite admission to a chiropractic program was found. Optimal weighting and adequate sampling are important for obtaining optimal generalizability. Limitations and suggestions for future research are discussed. (J Chiropr Educ 2009;23(1):8–16)

Key Indexing Terms: Education; Educational Measurement; College Admission Test; School Admission Criteria

INTRODUCTION

Admission and Attrition

A search of the websites of the 15 Council of Chiropractic Education (CCE) accredited chiropractic colleges in the US performed in August 2008, revealed that at least 14 of these used one or more 'non-cognitive' admission criteria, typically: letters of recommendation, references, written essays and admission interviews. Similarly, the four CCE accredited European chiropractic programs all appeared to use at least an admission interview of applicants as part of their admission procedure. The perceived need for using such admission tools in chiropractic education to complement admission based purely on previous grades is understandable. While a grade point average traditionally used in student selection may provide some information regarding applicants' average academic skills and their diligence across a range of different sciences, it does not guarantee their motivation for, or their knowledge of, the program or the future profession chosen. The completion of a demanding program, in this case a 5 year MSc chiropractic program, is likely

to depend on a range of factors in addition to general education and study skills. Such factors include genuine subject interest, motivation, social situation, social skills and perseverance. Existing qualitative and quantitative research on attrition indicate that a substantial number of Danish university students might be wrongly matched with a program or profession from the outset - at the point of admission.[1–2]

Complementary Admission

In 2002 the Faculty of Health Sciences at the University of Southern Denmark developed an admission to complement selection based purely on grades, to try to sift out applicants that were most reflective and informed about the programs and future careers and who also displayed interpersonal skills relevant for careers in the health science programs delivered (medicine, public health, chiropractic, sports science). The supplementary battery of admission criteria for the chiropractic program presented here consisted of measures of levels of qualification, motivation, common knowledge, and admission interview performance. Few other test situations in education deserve to be designated 'high-stakes' more than an admission test. Therefore, the right balance between reliability validity, acceptability and feasibility of the selection process is paramount.[3]

Reliability in Admission

Reliability in admission is important because it reflects on the extent to which applicants can be ranked consistently. Unreliable selection is not just a harmless exercise - it is unethical, useless and a waste of institutional resources. In the end, unreliable selection is therefore a liability for all stakeholders of the program. For the educational researcher, reliability is also important because it places an upper limit on the subsequent maximum possible test validity, because the maximum test validity obtainable is the square root of its reliability coefficient.[4] Once test reliability is known, it is in fact possible to correct validity studies for unreliability,[5] thereby effectively increasing the signal by noise reduction. Validity studies in selection are also paramount, McManus[3] formulates it nicely, "All assessments in selection are implicit predictions of future behaviour of a candidate. If there is no correlation with those future behaviours then they are not useful, however much assessors may agree about them."

A search of the databases ERIC, psychINFO, and PubMed with the search terms admission/selection and chiropractic and reliability/generalizability did not reveal any studies on the reliability of admission to chiropractic programs for comparison. In fact, most research on student selection to health professions has been done in the field of medical education, where competition for places is typically very fierce. There is a preponderance of evidence in support of using previous academic achievements (eg A-levels, MCAT scores etc.) as selection criteria in medical education.[5] However, while previous academic achievement is the best known independent predictor of pre-graduate success in medical[5–8] and chiropractic education[9], even at its best, it is only of moderate strength. Unfortunately, the overall evidence on most other predictors (often termed 'non-cognitive' predictors) is comparatively scarce[5] and even less convincing.[3,8,10–14]

The admission interview is widely used in medical education[15] and it is probably the best examined of the non-cognitive selection tools.[10] While the admission interview appears to have at least high 'face validity' for testing non-cognitive skills, such as communication skills and interpersonal skills, its reliability is controversial in health science education.[8,15–17] Even less evidence exists on the reliability (and validity) of submitted written statements as selection tools such as our 'written motivation' or the 'essay' used at some US chiropractic programs.[8,11,18]

Only a handful of studies in education examined the reliability of admission criteria with generalizability theory, and most of these revolve around the admission interview in medicine.[15,16,19–23] Very little has been published on the generalizability of composite admission procedures in medical education.[21,22] Classical reliability measures conceptualize a measurement as the sum of a true score and undifferentiated error. Generalizability theory is an extension of classical test theory, which allows for disentanglement and estimation of multiple source of error variance within the same study.[24] The results may reveal important sources of error variance, which may be targeted specifically to improve overall test reliability. Subsequent mathematical modeling allows for reliability estimates of both actual test conditions and alternative test strategies tailored to specific settings and needs. This makes generalizability theory more flexible than classical test theory.[4,24–28]

The aim of this study was: to estimate the generalizability of the admission to the chiropractic program during the admission in spring 2007. The objectives were:

1. to estimate the contributions to variance in scores,
2. to estimate the generalizability of individual admission parameters used, and
3. to estimate the overall generalizability of the composite admission process to chiropractic.

METHODS

Participants

The sampling frame was all the eligible applicants to the course of chiropractic who participated in the admission test in May 2007 at the Faculty of Health Sciences, at the University of Southern Denmark. All applicants with a sufficient grade point average (GPA) for the chiropractic course were invited to participate in the admission test. Fifty students were to be selected eventually: ten based on the highest GPA only, and forty based on their composite admission test rank.

Ethics

The study was registered with the Regional Ethics Committee and the Danish Data Protection Agency and fulfilled their requirements before it was initiated.

Admission Variables

Participants were scored on four admission variables: qualification, motivation, common knowledge, and an admission interview.

The qualification score was derived from the application form. Applicants submitted a standard national application form which contained specific questions developed nationally by Coordinated Application to be used in admission to all higher education in Denmark. A scoring manual was developed by the admission group of the Faculty of Health Sciences at the University of Southern Denmark. Scores were assigned for: Relevance and quantity of previous work experiences (0–35 points), past educational qualifications (0–35 points), foreign exchange experiences (0–10 points) and organizational/voluntary work (0–20 points). Highest scores were assigned to: Jobs involving care, qualifications in health sciences, foreign exchange experiences

involving care/voluntary work, and leadership experience. An overall qualification score between 0–100 points was assigned to each applicant by summing up scores in each category.

The motivation measure was based on a written statement in the essay format to assess: written communication skills, knowledge of the chosen course and profession, reflections on past experiences, reflections on choice of study, and future employment plans. The written motivation was prepared on-site as part of the admission test. Each of the five sub-domains was tentatively assigned a plus or a minus score at first only to guide an overall score, which was one direct score between 0–100 on a global rating scale.

Common knowledge was assessed with a 60 question/15 minutes multiple choice test, consisting of a wide variety of content sub-domains, eg: biology, physics, arts, news, music, health, politics etc. The format was 'one best answer'. The number of correct answers in the common knowledge test was converted to a 0–100 percentage score. The test was developed by a chief psychologist from the Section for Selection, a part of the Institute for Military Psychology in the Danish Army. The common knowledge test was administered on the admission test day.

The admission interview was a 25 minute semi-structured interview to assess: subject interest, expectations, maturity for age, social skills, stress tolerance, empathy, and general interview behavior. Each of these domains was given a tentative score of 1–5, only to guide the interviewer to an overall score, which was one direct score between 0–100 on a global rating scale. A list of 68 appropriate key questions was available, but interviewers were free to supplement these with their own questions where necessary to uncover relevant information on the domains to be assessed only. A score between 0–100 on a global rating scale was given. The admission interview took place on the admission test day.

The 1–100 global rating scales used for the written motivation and the interview were subdivided into 11 numbered points, eg 0, 10, 20. . . . 100, mainly to ensure a self-explanatory midpoint. No testing of the scales has currently been undertaken. Each admission variable was, for the purpose of this study scored by two random, independent raters on separate marking sheets. The common knowledge questionnaires were scanned and scored electronically on two separate occasions giving rise to two independent ratings of the same performance. The

raters of qualification and motivation were experienced faculty members, and they were blinded with regards to scores other than their own. Interview panels consisted of one faculty and one student interviewer, and each panel interviewed around 6–7 applicants. All interview raters were given thorough instructions on good habits for independent scoring on at least two occasions both in writing and verbally. To prevent bias, they were asked to avoid discussing scores before scoring, to keep a neutral body language until scoring was complete, and not to change the first given score. An afternoon training session was held for interviewers.

Analysis

For quality control purposes, paper data was converted to electronic data twice by two different operators. The two entries were compared with each other once and with the original scores on the marking sheets once. The disattenuated correlation matrix presented in the results was estimated with STATA 9.2 (StataCorp LP, College Station, TX) using the Best Linear Unbiased Predictor (BLUP) method. Variance components for a mixed model were estimated with the restricted maximum likelihood (REML) method. This method was used because the design was incomplete and unbalanced. The data can be described as 'naturalistic'[29] and as such contained both crossed and nested situations between applicants and raters. However, it was possible to estimate three variance components using the REML method: Firstly, the applicant variance (σ^2_p) or the applicant effect, which is the variance in scores that can be attributed to applicant differences, i.e. poor vs. good performers. Secondly, the rater variance (σ^2_r) or the rater effect, which is the variance in scores due to difference in rater stringency or leniency. And finally, a residual ($\sigma^2_{pr,e}$), which is the variance attributable to the applicant-rater interaction plus random error. Analysis was performed using STATA 9.2, which supplied the standard errors (SE), the confidence intervals (CI), and the co-variances and correlations used to calculate G-coefficients.

The generalizability coefficient (G), a measure of the reliability of the obtained rank order of an applicant, was calculated for each individual admission variable using the formula:

$$G = \sigma^2 p/(\sigma^2 p + \sigma^2 pr,e/n_r),\qquad(1)$$

where n_r is the number of raters used in the discipline. Generalizations were made to a universe of a random rater and a fixed admission variable on this occasion. The Decision studies for composite G-coefficients for the final composite selection processes were calculated with mGENOVA version 2.1 (Robert L. Brennan, Iowa Testing Programs, University of Iowa) by direct input of the correlations, variances, and co-variances for each admission variable estimated with STATA 9.2. In mGENOVA, composite p, r and pr,e variance components were found by summing the weighted elements in the respective variance-covariance matrices. From this, a composite generalizability coefficient, G_c, was derived.[25] The formula for estimating G_c, coefficients with mGENOVA was:

$$G_c = \sigma^2 p_c/(\sigma^2 p_c + \sigma^2 pr,e_c),\qquad(2)$$

where $_c$ denotes that composite variance components were used. The composite index of dependability, Φ_c, a measure of the composite reliability of the absolute score of an applicant, was calculated in a similar manner with the formula:

$$\Phi_c = \sigma^2 p_c/(\sigma^2 p_c + \sigma^2 r_c + \sigma^2 pr,e_c)\qquad(3)$$

The weights assigned to each variable in the composite G for the 2007 admission were based on administrative stakeholders' intuition. Alternative test strategies presented here as illustrative examples were found by performing alternative multivariate decision studies (D-studies) based on the estimated variance components from the 2007 admission. The alternative test strategies were found by weighting those admission parameters with the higher G more and assigning more raters to the admission parameters with low G.

RESULTS

105 invited applicants participated in the chiropractic admission test, yielding complete data from all 105 participants for analysis. Of these, 60/105 (57.1%) were women and 45/105 (42.9%) were men. Their average age was 21.3 years (SD = 2.5 years). The applicant nationalities were: 62/105 (59.0%) Danish, 24/105 (22.9%) Norwegian, 14/105 (13.3%) Swedish, and 5/105 (4.8%) were of other nationalities. Of the 105 applicants, 64/105 (61.0%), had the chiropractic program as their first priority on the application form.

The disattenuated correlation coefficients between admission variables were low-moderate (Table 1).

Table 1. Disattenuated correlation matrix[a] for admission variables to chiropractic in 2007 (N = 105)

	Knowledge	Interview	Qualification	Motivation
Knowledge	1.0			
Interview	0.19*	1.0		
Qualification	0.18	0.23*	1.0	
Motivation	−0.06	0.37***	0.08	1.0

[a]the correlation between variables after removal of measurement error. *p ≤ 0.05. **p ≤ 0.01. ***p ≤ 0.001.

Table 2. Variance components for admission variables to chiropractic in 2007

Effect	p				R				p x r,e			
Variable	σ^2	SE	%	df	σ^2	SE	%	df	σ^2	SE	%	df
Knowledge	152.81	21.09	100.00	104	N/A	N/A	N/A	1	0.00	0.00	0.00	104
Interview	348.06	55.36	78.06	104	5.70	6.65	1.28	35	92.04	13.72	20.65	70
Qualification	223.23	36.50	70.72	104	18.17	16.60	5.76	3	74.25	10.36	23.52	102
Motivation	273.22	59.62	49.25	104	13.40	18.72	2.42	3	268.08	37.25	48.33	102

The p effect is the variance attributable to differences in applicants. The r effect is the variance attributable to differences in rater stringency/leniency. The pr,e effect (residual) is the variance attributable to applicant and rater interaction plus random error, which in this case equals the relative error. σ^2 = estimated variance component value, SE = standard error, % = the percentage of total variance for the admission variable, df = degrees of freedom. N/A = not applicable.

The estimated variance components and their relative contribution to the total variance in scores are displayed for each admission parameter in Table 2. Even though the interview, the motivation and qualification variables had the largest values of applicant variance they also had the largest residual values, which in this case equal the relative error variance (Table 2). The written motivation in particular had a large residual contributing almost as much to the total variance in motivation scores as the applicant did. Differences in rater stringency (rater effect) generally contributed relatively little (≤5.76%) to the total variance in scores (Table 2) for all four admission variables.

The relationship between the applicant variance (the true score estimate) and the residual (relative error variance) for each admission variable is reflected in the estimated generalizability coefficients (G) in Table 3. The common knowledge test would most likely have reached G-coefficients of at least 0.90 even if only one rating had been used (Table 3). The admission interview and qualification needed at least 3 raters for a G greater than 0.90 (Table 3). The written motivation would most likely not have reached that level of reliability even with 3 ratings per applicant (Table 3).

In the actual admission to chiropractic in 2007, the intuitive weighting and number of raters used for each admission variable were as outlined in Table 4. The composite generalizability of an obtained applicant rank, G_c, in the 2007 admission was found to be 0.80 (Table 4). The reliability of the final composite absolute admission score of an applicant, Φ_c, was only slightly lower (0.78, Table 4). Alternative 1 and 2 in Table 4 are illustrative examples of alternative test strategies, based on the estimated variance components of the 2007 admission reported in Table 2. They are only two examples of many other possible alternatives.

DISCUSSION

Overall, good generalizability was found for the composite admission procedure of a chiropractic program, which consisted of four individual admission parameters.

The disattenuated correlation coefficients between admission variables were low to moderate, indicating acceptable discriminant validity of individual parameters (Table 1).

Table 3. Generalizability of individual admission elements for chiropractic in 2007

	$N_{r=1}$		$N_{r=2}$		$N_{r=3}$	
	G	[95% CI]	G	[95% CI]	G	[95% CI]
Knowledge	1.00	[1.00–1.00]	1.00	[1.00–1.00]	1.00	[1.00–1.00]
Interview	0.79	[0.72–0.87]	0.88	[0.84–0.93]	0.92	[0.89–0.95]
Qualification	0.75	[0.67–0.83]	0.86	[0.80–0.91]	0.90	[0.86–0.94]
Motivation	0.50	[0.36–0.65]	0.67	[0.54–0.80]	0.75	[0.65–0.86]

N_r = the number of raters/ratings. G is the reliability coefficient of the obtained rank order of an applicant. $G = \sigma^2 p/(\sigma^2 p + \sigma^2 pr,e/n_r)$, using the estimated variance components (σ^2) for the p and the pr,e effects reported in table 2.

Table 4. Composite generalizability coefficients for admission to chiropractic 2007 and alternatives

	Admission 2007	Alternative 1	Alternative 2
Knowledge, n_r (wt).	1 (0.20)	1 (0.25)	1 (0.20)
Interview, n_r (wt).	2 (0.25)	2 (0.50)	2 (0.25)
Qualification, n_r (wt).	1 (0.35)	1 (0.20)	3 (0.35)
Motivation, n_r (wt).	1 (0.20)	1 (0.05)	3 (0.20)
$\sigma^2 p_c$ (SD)	91.70 (9.58)	130.38 (11.42)	91.70 (9.58)
$\sigma^2(\delta)_c$ (SD)	22.70 (4.76)	15.15 (3.89)	9.48 (3.08)
$\sigma^2(\Delta)_c$ (SD)	25.64 (5.06)	16.62 (4.08)	10.58 (3.25)
G_c	0.80	0.90	0.91
Φ_c	0.78	0.89	0.90

N_r = the number of raters/ratings, wt = the weights given to individual parameters, $\sigma^2 p_c$ is the composite applicant variance, $\sigma^2(\delta)_c = \sigma^2 pr,e_c$ and is the composite relative error variance, $\sigma^2(\Delta)_c = \sigma^2 r_c + \sigma^2 pr,e_c$ and is the composite absolute error variance. G_c is the composite generalizability coefficient, a measure of the reliability of a composite applicant rank. $G_c = \sigma^2 p_c/(\sigma^2 p_c + \sigma^2(\delta)_c)$. Φ_c is the composite Index of Dependability, a measure of the reliability of the composite absolute score assigned to an applicant. $\Phi_c = \sigma^2 p_c/(\sigma^2 p_c + \sigma^2(\Delta)_c)$.

The variance component 'rater effect' (Table 2) is not applicable (N/A) for the common knowledge test, because no human rater was involved, rather this multiple choice test was scored electronically on two separate occasions. It was included in this study, because it was a weighted component in the composite admission, and as such contributed with applicant variance to the final composite admission rank.

Generalizability of Individual Variables

Error is naturally omnipresent in assessments. A proactive and responsible approach to this condition —particularly in 'high stakes' situations - is to estimate the size and influence of error, and to curb it most efficiently within the frame of what is feasible in a given setting. Error variance from many sources (eg rater, item, occasion, other) is most likely the reality in any type of assessment—not just in student selection - but it can be diminished by adequate sampling of the facets contributing with error variance.

Qualification

Good generalizability (G = 0.75, n_r = 1) was found for our qualification variable (Table 3) with its four different content sub-domains. Oosterveld and ten Cate also investigated the generalizability of application form information for comparison. Their applicants were also scored on four domains but the format was of the short answer type whereas ours was simply scoring the submitted evidence of previous qualifications by means of a relatively structured scoring manual. They found poorer generalizability (G = 0.28–0.37, n_r = 1) for their type of application form variable.[21]

Written Motivation

Poor-moderate generalizability was found for the written motivation (G = 0.50, $n_{r=1}$, Table 3) which is in concordance with the existing literature on auto-biographical submissions in medical edu-

cation.[8,11,18,30] Salvatori reported inter-rater reliability coefficients for written submissions in admissions to health educations to be between 0.15–0.59.[8] Nayer and Howe found high overall test generalizability (G = 0.76–0.78) for an off-site written personal statement used in admission to physiotherapy. Broad sampling of content (17 item questionnaire) and raters (n_r = 3) as well as horizontal scoring of each item, is likely to have contributed positively to such a result.[20]

Admission Interview

Very good generalizability was found for the admission interview (G = 0.88, n_r = 2, Table 3). The generalizability of admission interviews have been examined to some extent within medical education. This research reveal very varied results with coefficients ranging from 0.27–0.86.[15,16,19,23,31,32] The strength of generalizability studies is the flexibility of their designs. The trade off is, that comparison with other research results is no longer straightforward. Suffice to say, that the G for our admission interview was in the high range compared to previous results. However, both item and occasion may have been acting as hidden facets to confound our results. Some researchers found error variance stemming from an item effect (difference in question difficulty) and its interactions to be important for admission interviews.[15,16,23,31,32] Interestingly, Kreiter et al[32] found that moving from a fully structured interview using the same questions for each applicant (item as a fixed facet) to a less structured format with random questions from a question bank for each applicant (item as a random facet), was not particularly influential on the generalizability obtained. It is well known that that students' performance in assessments generally is content specific, ie dependent on the topic tested.[33] Kreiter et al[16] managed to identify error variance from an occasion effect in an admission interview. So, it would probably be wise to include an occasion facet in future G-study designs if at all possible or at least view occasion as a potential hidden facet in the interpretation of results, if it is not included.

Composite Generalizability

We found good generalizability (G_c = 0.80) for the composite selection procedure to the chiropractic program (Table 4). The equivalent composite generalizability coefficient of the composite admission procedure for 307 medical applicants at the same

Faculty in 2007, which was a composite of qualification, common knowledge and interview scores only, was estimated to be 0.82.[22] One other published study reported a composite admission procedure to medical school and found moderate-good generalizability (G = 0.72–0.74) combining the three admission variables: application form information, admission interview and a study sample assessment procedure.[21]

Minimum reliability coefficients of 0.90 have been suggested for high stakes test situations.[27] In the case of admission to chiropractic in 2007, simply changing the intuitive weighting of individual admission variables in the composite to one based on evidence, a relatively simple and purely administrative task, could have secured a G_c of 0.90 (Table 4, alternative 1). Had the weighting used in 2007 been important for content validity reasons, a G_c of at least 0.90 could also have been reached by using more raters for the variables with the lowest G (Table 4, alternative 2). The latter alternative may of course affect the test feasibility negatively. Many other alternative test strategies combining both evidence based weighting and use of raters are of course also possible.

Limitations

The generalizability coefficients estimated in this study is based on a single administration of a test form similar to measures of internal consistency in classical test theory. Lower coefficients would be expected if elements of test stability (same form of a test on two different occasions) or stability and equivalence (two different forms on two different occasions) had been incorporated into the generalizability coefficient.[25] So, one limitation of this study is the confinement of the universe of generalization, in this case to: a fixed universe of admission criteria employed, and on one fixed occasion.

Another main limitation of this and similar studies typically occur by failing to disentangle important effects contributing to error variance (eg items, occasions, other), which in turn leads to confounding of results. Unfortunately, replication of an authentic admission tests on separate occasions, which would allow for the disentanglement of an occasion effect, is rarely feasible and would probably be unacceptable to many stakeholders. Likewise, items are not always readily disentangled from all of the admission parameters used, as was the case for our written motivation and our semi-structured admission interview. Effects which are not disentangled or identified

may act as 'hidden facets'. The variance associated with a hidden fixed effect will be confounded with the variance components actually estimated, and result in deflated error variance and inflated applicant effect and generalizability. In contrast, the variance associated with a hidden random facet may either deflate or inflate the estimated generalizability.[25] Future research on generalizability of admission procedures and other assessments in higher education should therefore try to disentangle as many important sources of variance simultaneously as feasible.

CONCLUSION

We found good generalizability for a composite admission procedure to a chiropractic program consisting of four individual admission parameters. Some confounding of results due to hidden facets can not be ruled out. Optimal weighting and adequate sampling are important for obtaining good generalizability.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

1. The Danish Ministry of Education. Facts and figures 2007. Key figures in Education 2007. Copenhagen, Denmark: The Danish Ministry of Education; 2008.
2. Studenterrådet ved Aarhus Universitet. Frafald og studiemiljø. Århus, Denmark: Studenterrådet ved Aarhus Universitet; 2000.
3. McManus, IC. Student Selection. In: Dent JA, Harden RM, editors. A practical guide for medical teachers. Edinburgh: Elsevier; 2005. p. 364–73.
4. Streiner DL, Norman GR. 2003. Health measurement scales: A practical guide to their development and use. Oxford: Oxford University Press; 2003.
5. Ferguson E, James D, Madeley L. Factors associated with success in medical school: systematic review of the literature. BMJ 2002 Apr;324:952–7.
6. Christensen J, Juul S. [Effects of admission qualifications on medical education]. Ugeskr Laeger 1999 Jan;161(3):257–62.
7. McManus IC, Smithers E, Partridge P, Keeling A, Fleming PR. A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. BMJ 2003;327:139–42.
8. Salvatori P. Reliability and validity of admissions tools used to select students for the health professions. Adv Health Sci Educ Theory Pract 2001;6:159–75.
9. Green BN, Johnson CD, McCarthy K. Predicting academic success in the first year of chiropractic college. J Manipulative Physiol Ther 2003;26:40–6.
10. Goho J, Blackman A. The effectiveness of academic admission interviews: an exploratory meta-analysis. Med Teach 2006;28:335–40.
11. Kulatunga-Moruzi C, Norman GR. Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. Teach Learn Med 2002;14:34–42.
12. McManus IC, Iqbal S, Chandrarajan A, Ferguson E, Leaviss J. Unhappiness and dissatisfaction in doctors cannot be predicted by selectors from medical school application forms: a prospective, longitudinal study. BMC Med Educ 2005;5:38.
13. Lawson, DM, Till H. Predictors of performance of students from the Canadian Memorial College on the licensure examinations of the Canadian Chiropractic Examination Board. J Manipulative Physiol Ther 2006; 29:566–9.
14. Benbassat, J., and Baumal R. Uncertainties in the selection of applicants for medical school. Adv Health Sci Educ Theory Pract 2007;12:509–21.
15. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ 2004;38:314–26.
16. Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. Adv Health Sci Educ Theory Pract 2004;9:147–59.
17. Mrozek JP, Till H, Taylor-Vaisey AL, Wickes D. Research in chiropractic education: An update. J Manipulative Physiol Ther 2006;29:762–73.
18. Dore KL, Hanson M, Reiter HI, Blanchard M, Deeth K, Eva KW. Medical school admissions: enhancing the reliability and validity of an autobiographical screening tool. Acad Med 2006;81(10 Suppl):S70–3.
19. Harasym PH, Woloschuk W, Mandin H, Brundin-Mather R. Reliability and validity of interviewers' judgments of medical school candidates. Acad Med 1996;71(1 Suppl):S40–2.
20. Nayer M, Howe J. The Reliability of a personal statement as a selection tool in the admission process. In: Scherpbier AJJA RvdSA van der Vleuten CPM, editors. Advances in Medical Education. Dordrecht: Kluwer Academic Publisher; 1997. p. 465–78.
21. Oosterveld P, ten Cate O. Generalizability of a study sample assessment procedure for entrance selection for medical school. Med Teach 2004;26:635–9.
22. O'Neill LD, Korsholm L, Wallstedt B, Eika B, Hartvigsen J. Generalizability of a composite student selection program. Med Educ In press 2008.
23. Roberts C, Walton M, Rothnie I, Crossley J, Lyon P, Kumar K, Tiller D. Factors affecting the utility of the multiple mini-interview in selecting candidates

for graduate-entry medical school. Med Educ 2008;42: 396–404.

24. Boulet JR. In: B.S. Everitt BS, Howell DC, editors. Generalizability theory: Basics. Vol. 2 of Encyclopedia of statistics in behavioral science. Chichester: John Wiley & Sons, Ltd.; 2005. p. 704–11.
25. Brennan RL. Generalizability theory. New York: Springer-Verlag; 2001.
26. Crossley J, Davies H, Humphris G, Jolly B. General-isability: a key to unlock professional assessment. Med Educ 2002;36:972–8.
27. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006–12.
28. Shavelson RJ, Webb NM. Generalizability theory. A Primer. Newbury Park: Sage; 1991.
29. Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, et al. 'I'm pickin' up good regressions':

the governance of generalisability analyses. Med Educ 2007;41:926–34.

30. Hanson MD, Dore KL, Reiter HI, Eva KW. Medical school admissions: revisiting the veracity and indepen-dence of completion of an autobiographical screening tool. Acad Med 2007;82(10 Suppl):S8–11.
31. Arce-Ferrer AJ, Castillo IB. Investigating postgraduate college admission interviews: generalizability theory reliability and incremental predictive validity. J Hisp High Educ 2007;6:118–34.
32. Kreiter CD, Solow C, Brennan RL, Yin P, Ferguson K, Huebner K. Examining the influence of using same versus different questions on the reliability of the medical school preadmission interview. Teach Learn Med 2006; 18:4–8.
33. Wass V, der Vleuten CV, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357:945–9.