



Published in final edited form as:

Acad Med. 2010 January ; 85(1): 169–176. doi:10.1097/ACM.0b013e3181c464c0.

Utility of the AAMC's Graduation Questionnaire to Study Behavioral and Social Sciences Domains in Undergraduate Medical Education

Patricia A. Carney, PhD [Professor],

Department of Family Medicine and Department of Public Health and Preventive Medicine, Oregon Health and Science University, Portland, Oregon

Rebecca Rdesinski [Research associate],

Department of Family Medicine, Oregon Health and Science University, Portland, Oregon

Arthur E. Blank, PhD [Associate professor],

Department of Family and Social Medicine and Department of Epidemiology and Populations Health, Albert Einstein College of Medicine, Bronx, New York

Mark Graham, PhD [Assistant professor],

Center for Education Research and Evaluation, Columbia University College of Physicians and Surgeons, New York, New York

Paul Wimmers, PhD [Assistant professor],

Center for Education Development and Research, David Geffen School of Medicine, University of California, Los Angeles, California

H. Carrie Chen, MD [Associate clinical professor],

Department of Pediatrics, University of California School of Medicine, San Francisco, California

Britta Thompson, PhD [Assistant professor],

Office of Undergraduate Medical Education, BCM, Baylor College of Medicine, Houston, Texas

Stacey A. Jackson [Associate director],

Dean's Office of Medical Education and Curricular Affairs, Indiana University School of Medicine, Indianapolis, Indiana

Julie Foertsch, PhD [Associate scientist], and

Department of Innovations in Medical Education, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

David Hollar, PhD [Research assistant professor]

Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina

Abstract

Correspondence should be addressed to Dr. Carney, Department of Family Medicine, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd. Mail Code: FM, Portland, OR 97239; telephone: (503) 494-9049; fax: (503) 494-2746; carneyp@ohsu.edu.

Other disclosures: None.

Ethical approval: All nine NIH-awardee medical schools underwent IRB review at their respective institutions for participation in this study and sharing of their CQ data, and all received exemptions from their IRBs.

Previous presentations: A poster was presented at the Association of American Medical Colleges annual meeting, November 2–7, 2007, Washington, DC.

Purpose—The Institute of Medicine (IOM) report on social and behavioral sciences (SBS) indicated that 50% of morbidity and mortality in the United States is associated with SBS factors, which the report also found were inadequately taught in medical school. A multischool collaborative explored whether the Association of American Medical Colleges Graduation Questionnaire (GQ) could be used to study changes in the six SBS domains identified in the IOM report.

Method—A content analysis conducted with the GQ identified 30 SBS variables, which were narrowed to 24 using a modified Delphi approach. Summary data were pooled from nine medical schools for 2006 and 2007, representing 1,126 students. Data were generated on students' perceptions of curricular experiences, attitudes related to SBS curricula, and confidence with relevant clinical knowledge and skills. The authors determined the sample sizes required for various effect sizes to assess the utility of the GQ.

Results—The 24 variables were classified into five of six IOM domains representing a total of nine analytic categories with cumulative scale means ranging from 60.8 to 93.4. Taking into account the correlations among measures over time, and assuming a two-sided test, 80% power, alpha at .05, and standard deviation of 4.1, the authors found that 34 medical schools would be required for inclusion to attain an estimated effect size of 0.50 (50%). With a sample size of nine schools, the ability to detect changes would require a very high effect size of 107%.

Conclusions—Detecting SBS changes associated with curricular innovations would require a large collaborative of medical schools. Using a national measure (the GQ) to assess curricular innovations in most areas of SBS is possible if enough medical schools were involved in such an effort.

The Institute of Medicine (IOM) released a report in 2004 summarizing how undergraduate medical school curricula should be enhanced to address critical health issues faced in the United States.¹ One major finding was that approximately half of all causes of mortality in the United States are linked to social and behavioral factors, such as smoking, diet, alcohol, sedentary lifestyle, and accidents.² It is generally recognized that biomedical research cannot address these issues alone. Rather, it will likely take the work of multidisciplinary collaborations to study the interactions between biology and behavior. Less than 5% of the more than \$2 trillion spent on health care annually in the United States is devoted to reducing behavioral and social risk factors.^{3,4} The IOM also found that the curriculum in most U.S. medical schools does not provide sufficient training about these behavioral and social risk factors, despite the fact that significant mortality and morbidity reductions could be realized.¹

In response to the IOM report, the National Institutes of Health (NIH) awarded grants to nine medical schools* to develop, pilot, and disseminate social and behavioral sciences (SBS)-modified curricula across the six domains identified by the IOM: (1) Mind-Body Interactions in Health and Disease, (2) Patient Behavior, (3) Physician Role and Behavior, (4) Physician-Patient Interactions, (5) Social and Cultural Issues in Health Care, and (6) Health Policy and Economics. The collaborations among the schools and the curricular innovations that emerged are described in detail elsewhere.⁵ Briefly, the projects vary with respect to the focus of the interventions, but all nine medical schools are incorporating SBS content throughout all four years of medical school in both the preclinical and clinical curricula. Of note is that the IOM recommended *revised curriculum integration* rather than the *development of new courses*. Examples of the curricular changes implemented include incorporation of biopsychosocial approaches that stress holistic, culturally sensitive, and interactive approaches to patient care,

* Albert Einstein College of Medicine; Baylor College of Medicine; Columbia University College of Physicians and Surgeons; David Geffen School of Medicine at the University of California, Los Angeles; Indiana University School of Medicine; Oregon Health and Science University; University of California, San Francisco, School of Medicine; University of North Carolina School of Medicine; and the University of Wisconsin School of Medicine and Public Health.

development of student empathy, communication, and teamwork skills with a particular focus on patient safety,^{6–8} and promotion of lifelong habits of self-directed learning and self-care.

Approximately 6,100 medical students will be affected by curricular innovations implemented during the five years of the collaborative.⁵ Individual institutions in the collaborative are using a variety of evaluative methods (e.g., curricular mapping, logic modeling, qualitative and quantitative assessments) to evaluate the effectiveness of these curricular innovations, and we believe that the results from these evaluations will help promote the dissemination of effective components to other medical schools. To optimize evaluative interactions within the collaborative, one school, Oregon Health and Science University (OHSU), submitted an administrative supplement to create an evaluation core. This core, overseen by one of us (P.A.C.), would establish a database of various outcomes from the nine schools, increase sample size from one school to nine, and allow for more robust analyses.

In the present report, we describe how we explored the utility of using existing multiinstitutional data from the Association of American Medical Colleges (AAMC) Graduation Questionnaire (GQ) to detect changes in students' scores related to curricular innovations over time. The GQ is a Web-based survey administered annually to all students graduating from allopathic medical schools in the United States. The GQ allows the collection of information on student characteristics as well as evaluations of the medical school experience. The summary data from this survey are available to all medical schools in the United States. We hypothesized that this standardized survey instrument could inform the development of relevant elective activities, management of curricular time, improvements in confidence with clinical knowledge and skills, and attitudes toward medicine and the medical profession. We undertook a modified Delphi approach⁹ to come to agreement on which variables could potentially capture curricular change and whether our collaborative efforts would provide enough statistical power to do so. Although we acknowledge that the GQ was not designed for this specific purpose, if the instrument could be used in the way we hypothesize, then all medical schools in the United States might benefit from its use to assess these important areas of medical training that the IOM identifies.

Method

OHSU's IRB review included the request to receive anonymous study data at a central location for analyses. Participating schools' identifiers were replaced with a study code for data in all pooled analyses. Starting in July 2006, we conducted a content analysis¹⁰ of the GQ to identify questions that could be used to assess elements of the IOM's reported SBS domains. The GQ questions were presented at a meeting that involved all principal investigators and evaluators from each of the nine schools. The selected question set (see the Appendix) was reviewed and reduced on the basis of relevance to planned curricular activities.

We next created a classification matrix that we sent to the evaluators at each participating medical school. Investigators classified each GQ question according to the six IOM domains, with the understanding that a question may map to more than one domain. This matrix was then returned to the evaluation core, and data were pooled to examine agreement among the raters. A modified Delphi approach⁹ was used to come to agreement on classification of study variables via teleconferencing. Significant empirical variations were initially evident in how evaluators mapped the GQ questions to the IOM domains, with overlap into more than one domain area being the greatest source of variation. Although it is likely that GQ questions will correlate with each other and across domains, for purposes of simplicity during this exploratory, nonquantitative phase, we constrained questions to single domains (i.e., factors) with presumably the greatest group agreement (in lieu of eigenvalues).

During the modified Delphi process, participants reached consensus on using a rate of 75% or higher (essentially seven out of nine schools, or 78% in agreement) as an automatic criterion for agreement, which occurred for 87.5% (21/24) of variables. Questions with less than 12% difference in overlapping domains (e.g., one domain with 78% and another with 89% agreement; $n = 3$) were discussed until consensus was reached to use the domain with the highest level of consensus. Using these approaches, we established that none of the GQ questions mapped to the Mind-Body Interactions in the Health and Disease domain. The final set of 24 questions mapped to five of the six IOM domains and are distributed among three question categories: (1) students' perceptions of the adequacy of appropriate curricular time, (2) level of confidence with clinical knowledge and skills, and (3) attitudes toward medicine and the profession. This distribution created nine matrix subtypes into which the GQ questions mapped. (see Table 1).

We pooled summary data from the 2006 and 2007 GQ from all nine participating medical schools in the collaborative, representing 1,126 students. The 2006–2007 project year was essentially a planning year, as it preceded any of the school's interventions, and was considered to serve as a baseline measurement period. Responses to many GQ questions were categorical. To normalize the data for presentation, we collapsed agreement response categories (e.g., agree and strongly agree) into one category and analyzed these responses as continuous variables with a potential range from 0% to 100% by considering any agreement as the numerator and the total number of responses as the denominator. Using this approach allowed us to calculate means and treat these variables as continuous. The mean frequency for *agree* was 42.4% (range: 11.0%–62.7%) of respondents; for *strongly agree*, it was 37.7% (range: 0.0%–77.6%). Similarly, we used the single category of *% appropriate curricular time* and treated it as a continuous variable. Whereas the other categories reflected inadequate curricular time or too much curricular time, both these categories reveal dissatisfaction with sufficiency in how curricular time is spent. The mean for inadequate time for all scores combined was 15.1% (range: 0.0%–36.9%), and the mean for too much curricular time was 4.6% (range: 0.0%–21.3%), indicating the most common perception of how time was spent was in the appropriate category.

Using SPSS (v15; Statistical Package for the Social Sciences, SPSS, Inc., Chicago, Illinois, 2007), we calculated means and standard deviations for the nine-school collaborative. Using the middle-ranked standard deviation of the nine matrix subtypes and the following effect sizes, 0.20, 0.50, 0.80, we then calculated the mean differences we could observe. Using those mean differences, a two-sided dependent-samples *t* test, 80% power, and alpha of .05, in conjunction with G*Power software,¹¹ we estimated the sample size required to observe each effect size. We treated each school as its own control during the baseline period. This approach assisted us in understanding how many schools would be needed to detect differences that could be attributed to curricular innovations. Additionally, using a two-sided test, 80% power, an alpha of .05, and a sample size of nine schools for each pre and post group, we estimated the effect size that we could detect using the nine-school collaborative.¹² Last, we evaluated how comparisons to national mean summary scores could assist in measuring change using the GQ.

All nine NIH-awardee medical schools underwent IRB review at their respective institutions for participation in this study and sharing of their CQ data, and all received exemptions from their IRBs.

Results

Characteristics of the medical students at participating medical schools, and response rates to the GQ, are presented in Table 2.¹³ Response rates ranged from 48% to 100%, with five of the nine schools having response rates >75% on all GQ items used in this study. Schools

represented both public and private institutions ($n = 6$ and 3 , respectively) and were primarily urban ($n = 8$). Institutional characteristics were removed from the table and summarized in general form in the legend to protect the confidentiality of schools.

Table 3 presents the means, standard deviations, and ranges of GQ responses according to matrix subtype (IOM domain and GQ question category). Mean scores of the nine participating schools ranged from a low of 60.8% (the percentage of students who agreed or strongly agreed with the question in the *physician role and behavior domain—attitude toward medicine and the profession* question category matrix subtype) to a high of 93.4% in the *physician role and behavior domain—confidence with clinical knowledge and skills* question category, indicating high scores for this matrix subtype. All of these scores from the collaborative were similar when compared with the mean scores of all U.S. medical schools combined (see Table 3).

Using the standard approach for calculating sample size required for a paired-sample t test,^{12,14} and the data we amassed that are shown in Table 3, we found that at least 34 medical schools would be needed to attain an estimated medium effect size of 0.50 (50%—see Table 4), which corresponds to a mean difference ranging from 0.6 to 4.4 (SDs: 1.2 to 8.8, respectively), to observe a statistically significant change. Using this same approach,^{11,12} but with a sample size of nine (the number of schools in the collaborative), we determined that we would need an effect size of 107% to detect statistically significant changes in our collaborative (see Table 4).

Similarly, consultation of Cohen's Table 2.3.5^{12(pp36–37)} for these specifications, 64 schools would be needed to achieve a power of 0.80 and a medium effect size difference of $d = 0.50$ standard deviation units. For the nine SBS consortium medical schools described in this study, Cohen's table estimates an enormous effect size difference of 1.40 standard deviation units for Power = 0.79, in contrast to the small effect size differences indicated in the GQ item results for our Table 3.

Discussion

From a multiinstitutional perspective, our study is among the first to examine the utility of using data from a national survey completed by all medical schools in the United States (i.e., the GQ) as a possible measure of the impact of curricular change in multischool studies. Whereas Newell¹⁴ used data from the GQ as part of an assessment of curricular change in geriatrics education, only a single institution was involved in that study. Pugnaire and colleagues¹⁵ used GQ data to assess the longitudinal stability of students' perceptions of the quality of their clerkships, and Marantz et al¹⁶ used GQ data to assess student satisfaction with curricular change; all of those studies involved single institutions. Although the GQ has also been used to assess changes in planned career settings¹⁷ and predictors of career choice,¹⁸ only one of those last two studies¹⁶ was a multiinstitutional study. Our study makes a substantial contribution to the literature on what would be involved in conducting a national assessment of changes in the SBS curriculum, as measured by the GQ. To our knowledge, no other group of educational researchers has analyzed the content of the GQ and classified it as we have done. We were hopeful that these efforts would allow our nine schools to use this measure to assess for changes over time.

The findings from our study suggest that for our learning collaborative to benefit from the GQ, using the assumptions we included in our analyses, the effect size associated with the planned innovations would have to be very large ($\geq 107\%$). Thus, our collaborative is likely too small, with its sample size of nine schools, to identify significant impacts of curricular change using this tool. We found that it would likely take a collaboration of more than 35 medical schools to observe modest changes (50%) in curricular impact over time, and that we would need a

sample size of 199 schools to detect a small effect size (20%), which is much greater than the number of medical schools that are currently enrolling medical students in the United States and Canada. Given that the impetus for our work was an important IOM extensive review, we believe it is important for the medical education community to understand the utility of existing measures and foundational aspects needed to study change over time.

The analytic challenge inherent in sample size calculations involves the balance of rigorous criteria, such as setting power at 0.80, using two-tailed tests, and setting appropriate alpha levels (.05 or .01 to account for multiple comparisons) with other parameters that can be expected to vary. In our case, we had a fixed sample size ($n = 9$) and available summary scores that allowed us to apply some assumptions regarding mean scores and standard deviations, leaving the effect size, or the difference we could attribute to curricular innovations, to finish the calculations. Though we found our small collaborative of nine schools could likely not find statistical differences associated with change, a much larger collaborative might have. Effect size differences for measurable change on attitude items in the GQ are likely to be small, thereby warranting a larger collaborative to measure such differences with adequate statistical power. We are encouraged by this work, and we hope others will undertake such studies so that larger collaboratives can form in meaningful ways.

This work illustrates that the GQ can enable the establishment of baseline measures that could be used to assess curricular changes in subsequent medical school classes across U.S. medical schools. Our study found that improvements could be detected for the majority of measures on the GQ that are related to SBS. Few areas would experience a ceiling effect, where statistical changes could not be detected because baseline scores were already high, given the effect sizes were large and enough medical schools participated in the analysis. The IOM report¹(p7) indicated that existing national databases provide inadequate information on behavioral and social science content, teaching techniques, and assessment methodologies. This lack of data impedes the ability to reach conclusions about the current state and adequacy of SBS instruction in U.S. medical schools. We have undertaken an important approach to remedy this situation. In fact, the IOM¹ report set the precedent for using the only available standardized questionnaire administered to all U.S. medical students, the GQ, by citing its use to measure medical student satisfaction with specific topics, most notably communication, cultural diversity, and socioeconomics.¹ Many articles have reported on the paucity of robust evaluation and educational research conducted in the United States.^{19–22} It is vitally important that this should change and for schools to collaborate more effectively toward a unified mission of enhancing medical education. It is important to note, however, that the GQ is not a completely standardized instrument in that some of the questions do change over time, which would affect the use of the tool in prospective studies. Retrospective analyses would need to take these changes into account.

With nine medical schools receiving NIH Behavioral and Social Sciences K07 awards, each institution had a particular vision and measurement approach. However, it became readily apparent that collaboration on common IOM domain themes could make the respective projects synergistically powerful. The OHSU administrative supplement created an evaluation core that would serve to develop plans and databases for pooled data and undertake relevant analyses, setting the stage for the nine-school NIH Behavioral and Social Sciences Consortium.⁵ Several schools have overlapping student outcome evaluation measures; however, finding a standardized measure that can be used across all institutions is beneficial. As these efforts progress during the nine schools' implementation phases, available baseline data, such as the GQ, serve as a logical starting point, and we plan to continue our assessments using this tool over time and, if possible, work with the NIH to expand the collaborative.

Using the GQ core questions to track the influence of the medical schools' interventions makes the assumption that this tool, which was not designed for this purpose, is sensitive enough to measure change. Our analytical approach essentially created a proxy measure for capturing change in the absence of an existing, psychometrically sound, measure. This creates both an opportunity and a challenge. The opportunity is to use the GQ as a proxy measure for trying to assess broad systemic changes in curriculum over time. However, as a proxy measure we realize that the GQ may not be sensitive enough to capture the benefits, or negative consequences, of very targeted curricular change. To the extent that curriculum change is medical-school-specific, new measures of curricular change will have to be developed that can more directly pick up the subtleties of those efforts. The modified Delphi approach we undertook allowed us to discuss this issue at length, and we hope those reading this report will understand and appreciate the challenges in this. In research, many measures are either adapted or used in ways that were different from their developmental purpose. This occurs because instrument development, testing, and administration are expensive undertakings. Our hope is that our collaborative would contribute to the dialogue about the use of national data sources.

With the exception of Mind-Body Interactions in the Health and Disease domain, the collaborative's modified Delphi approach achieved consensus on GQ questions corresponding to the remaining IOM domains: Patient Behavior, Physician Role and Behavior, Physician-Patient Interactions, Social and Cultural Issues in Healthcare, and Health Policy and Economics. Supplementing the GQ with more SBS questions in the Mind-Body Interactions in Health and Disease domain could potentially be important. Applications of these results will be useful for other medical institutions as they evaluate curricular change and as they address unique and common curriculum components. The approach could be useful for measuring Liaison Committee on Medical Education topic areas across institutions.

The IOM report notes that the domains it identifies are broad, and the report leaves each school to decide exactly what content to include. The local variations in curricula offered will be important to track. Given the possible variability of what will actually be implemented at each of the nine schools, the GQ core questions may be more or less sensitive to capture effects of the interventions. For example, a school that offers little content regarding cultural components of care may not expect to see a change in that component of the GQ core questions, or if a change is evident, it is possible the change has occurred for reasons other than the intervention. For the GQ core questions to be used to track the consequences of the medical school's interventions, and to reach its highest potential, it becomes critical to document how the actual curricula have been implemented in each of the schools. With that piece in place, and with enough medical schools contributing data, the use of the GQ would become a valuable tool in monitoring change.

The strength of our study is that nine schools collaborated and worked with an existing measurement tool to which all medical schools have access. The cost of the administrative supplement that supports the nine schools was minimal (\$40K annually), which could be contributed by other collaborative efforts at the institutional level if the value were explicitly recognized. The GQ has been administered to graduating medical students since 1978 to assist the AAMC and medical schools in priority setting and in program and policy development.²³ In addition, as indicated, several studies have been conducted using this tool,¹⁴⁻¹⁸ and many of those used less formal processes for coming to agreement on classifying study variables than we did.

An important limitation of the study is that we did not have access to data files containing individual responses from each medical student, which would have allowed us to work with the data without making assumptions required in the absence of individual-level rather than school-level summary data. Because of this, we have oversimplified the sample size

calculations we conducted, though we attempted to be as conservative as possible. We made multiple attempts to obtain individual schools' data from the AAMC directly via a request from the multischool core located at OHSU and via individual schools. Unfortunately, no request succeeded in obtaining a data file with individual responses. Although some schools were able to obtain GQ data in the past, in response to the present request, we were told that only summary data would be provided at the level of the school because of limited resources at the AAMC. If we had had access to individual-level data, we could have taken into account the clustering effects of students within specific institutions as well as expected effect sizes based on planned innovations in our calculations. This would also have allowed us to adjust for possible confounders, such as age and gender of students. However, the GQ is likely the best starting point for addressing these questions, given that an extensive, structured experimental design across the nine institutions is something that is beyond the funding capability of the K07. We hope that studies like the one we conducted will foster the use of national data in a way that can benefit all medical schools and students in training.

We also do not have explicit access to control schools, so we made some assumptions about historical cohort comparisons (using each school to serve as its own control) and assessments using mean national scores that would need to be tested using a more rigorous study design. Last, we considered only the category of appropriate curricular time spent in our analyses of some IOM domain areas, which did not take into account whether the remaining time was excessive or inadequate. The consequences of this are that although we make comparisons with some accuracy measures of appropriate time, we do not fully address excessive or inadequate time. However, our detailed exploration of this issue indicates that the most likely response was appropriate curricular time, with fewer than 20% of responses being inadequate curricular time (15.1%) or too much curricular time (4.6%); thus, our findings are not likely influenced to any significant extent by this approach.

Finally, we note that the response rates to the GQ range from 48% to 100% with a mean of 74%. There is the possibility of response bias affecting these findings, especially from the four schools with response rates below 75%. The schools use different techniques to encourage their students to complete the GQ, which included doing nothing, weekly e-mail reminders from administration, a student-run recruitment program, offering a nominal gift certificate to each student completing the survey, offering a nominal amount toward the Residency Match Day Fund, requiring completion of the survey to receive grades, and requiring completion for graduation (one school). These different attempts seemed to have influenced the students' response rates regarding survey completion. This may have entered bias into the study, even though the overall response rate was quite high (74%), and given the fact that some schools had more difficulty than others in attaining high response rates.

Conclusions

Using a national measure (the GQ) to assess changes associated with curricular enhancements in most areas of SBS is possible if enough medical schools were involved in such an effort. Medical schools should work collaboratively to take advantage of these nationally available data to assess their programs' ability to potentially improve the nation's health.

Acknowledgments

Funding/Support: This work was supported by the following NIH Institutes: NCI (K07 CA121457), NHLBI (K07 HL082628, K07 HL082629), NICHD (K07 HD051546, K07 HD051528), NIAMS (K07 AR53812, K07 HD051507), and NCAM (K07 AT003131, K07AT003346).

References

1. Institute of Medicine. Improving Medical Education: Enhancing the Behavioral and Social Science Content of Medical School Curricula. Washington, DC: National Academies Press; 2004.
2. McGinnis JM, Foege WH. Actual causes of death in the United States. *JAMA* 1993;270:2207–2212. [PubMed: 8411605]
3. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Prevalence Data. 1998 [Accessed 2008]. <http://ww2.cdc.gov/nccdphp/brfss/index.asp>[no longer available]
4. Health Care Financing and Administration, Department of Health and Human Services. HHS Fiscal Year 2000 Freedom of Information Annual Report. [Accessed September 30, 2009]. Available at: <http://www.hhs.gov/foia/reports/00anlrpt.html>
5. Hollar D, Satterfield JM, Carney PA, et al. The National Institutes of Health Social and Behavioral Science Consortium: An introduction and progress report on undergraduate medical education curricular innovations. *Ann Behav Sci Med Educ* 2007;13:60–68.
6. Hobgood C, Sherwood G, Frush K, et al. Teamwork training with nursing and medical students: Results of an inter-institutional, interdisciplinary collaboration. *Qual Saf Health Care*. In press.
7. Sherwood, G.; Frush, K.; Hollar, D. UNC–Duke GSK Interdisciplinary Team. Measuring the effects of high and low fidelity simulation on interdisciplinary teamwork. Proceedings of the Council for the Advancement of Nursing Science State of the Science Conference; Washington, DC: 2008.
8. Baker DP, Salas E, King H, Battles J, Barach P. The role of teamwork in the professional education of physicians: Current status and assessment recommendations. *Jt Comm J Qual Patient Saf* 2005;31:185–202. [PubMed: 15913126]
9. Helmer, O. The Systematic Use of Expert Judgment in Operations Research. Santa Monica, Calif: RAND Corporation; 1963.
10. Carney, PA.; Rdesinski, RE.; Blank, AE., et al. Efforts of a multi-institutional collaborative to study the impact of innovative curricula on behavioral and social sciences. Poster presented at: Association of American Medical Colleges Annual Meeting; November 2–7, 2007; Washington, DC.
11. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007;39:175–191. [PubMed: 17695343]
12. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
13. Barzansky B, Etzel SI. Medical schools in the United States, 2006–2007. *JAMA* 2007;298:1071–1077. [PubMed: 17785656]
14. Newell DB. Integrating geriatric content into a medical school curriculum: Description of a successful model. *Gerontol Geriatr Educ* 2004;25:15–32. [PubMed: 15778143]
15. Pugnaire MP, Purwono U, Zanetti ML, Carlin MM. Tracking the longitudinal stability of medical students' perceptions using the AAMC graduation questionnaire and serial evaluation surveys. *Acad Med* 2004;79(10 suppl):S32–S35. [PubMed: 15383383]
16. Marantz PR, Burton W, Steiner-Grossman P. Using the case-discussion method to teach epidemiology and biostatistics. *Acad Med* 2003;78:365–371. [PubMed: 12691965]
17. Jeffe DB, Andriole DA, Hageman HI, Whelan AJ. The changing paradigm of contemporary US allopathic medical school graduates' career paths: Analysis of the 1997–2004 national AAMC Graduation Questionnaire Database. *Acad Med* 2007;82:888–894. [PubMed: 17726402]
18. McAlister RP, Andriole DA, Rowland PA, Jeffe DB. Have predictors of obstetrics and gynecology career choice among contemporary US medical graduates changed over time? *Am J Obstet Gynecol* 2007;196:275.e1–275.e7. [PubMed: 17346553]
19. Baernstein A, Liss H, Carney PA, Elmore JG. Trends in study methods used in undergraduate medical education research, 1969–2007. *JAMA* 2007;298:1038–1045. [PubMed: 17785648]
20. Carney PA, Nierenberg DW, Pipas CF, Brooks WB, Stukel TA, Keller AM. Educational epidemiology: Applying population-based design and analytic approaches to study medical education. *JAMA* 2004;292:1044–1050. [PubMed: 15339895]

21. Dauphinee WD, Wood-Dauphinee S. The need for evidence in medical education: The development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. *Acad Med* 2004;79:925–930. [PubMed: 15383347]
22. Shavelson, RJ.; Towne, L., editors. *Scientific Research in Education*. Washington, DC: National Academy Press; 2002. Committee on Scientific Principles for Educational Research.
23. Association of American Medical Colleges Graduation Questionnaire. [Accessed March 20, 2009]. Available at: <http://www.aamc.org/data/gq>

Appendix 1

Social and Behavioral Science Questions on the 2006 and 2007 Association of American Medical Colleges Graduation Questionnaire

AAMC Graduation Questionnaire 2006

21. Based on your experiences, indicate whether you agree or disagree with the following statements about medicine and the profession: [Pages 38 and 40]

- Physicians will not receive the same respect from society in the future as they have in the past.
- Access to medical care continues to be a major problem in the United States.
- Everyone is entitled to receive adequate medical care regardless of his or her ability to pay.
- Physicians have an opportunity to exercise greater influence on health promotion and disease prevention.
- Cure of disease is the most important purpose of medicine.
- Relief of patient suffering is the most important pursuit of medicine.

AAMC Graduation Questionnaire 2007

10. Indicate the activities you will have participated in during medical school on an elective or volunteer (not required) basis: [Page 15]

- Global health experiences
- Delivering health services to underserved populations at a clinical site
- Providing health education (e.g., HIV/AIDS education, breast cancer awareness, smoking cessation)
- Experience related to minority health disparities
- Experience related to cultural awareness or cultural competence
- Worked on a project with a community based multicultural group

11. Do you believe that the time devoted to your instruction in the following areas was inadequate, appropriate, or excessive: [Pages 16, 17, 18, 20, and 24]

- Clinical Decision Making and Clinical Care
 - Patient interviewing skills
 - Physician-patients communication skills
 - Physician-physician communication skills
 - Teamwork with other health professionals
 - Ethical decision making
- Population Based Medicine
 - Culturally appropriate care for diverse populations
 - Health and health care disparities
 - Health determinants
- Other Medical Topics
 - Family/domestic violence
 - Biomedical ethics
 - Professionalism

15. Indicate your level of agreement with the following statements: I am confident that I have the appropriate knowledge and skills to: [Page 28 & 29] *

Communication Skills

Discuss a prescription error I made with a patient.

Provide safe sex counseling to a patient whose sexual orientation differs from mine.

Discuss treatment options with a patient with terminal illness.

Initiate discussion of DNR orders with a patient or family member.

Negotiate with a patient who is requesting unnecessary tests or procedures.

16. Indicate whether you agree or disagree with the following statements about your preparedness for beginning a residency program: [Page 31] *

I have the ethical and professional values that are expected of the profession

22. Based on your experiences, indicate whether you agree or disagree with the following statements about medicine and the profession: [Page 42] *

Physicians will not receive the same respect from society in the future as they have in the past

SUPP 16. Based on your experiences, indicate whether you agree or disagree with the following statements: [Page 53]

I was appropriately trained to care for individuals from racial and ethnic backgrounds different from my own. *

*The GQ question referred to "statements," but in this Appendix, we have listed only the statement that we used in our study.

Table 1

Number of Questions on the 2006 and 2007 AAMC Graduation Questionnaire (GQ) That Mapped to the Institute of Medicine's Social and Behavioral Science (SBS) Domains, by Matrix Subtype*

| GQ question category | SBS domain and number of GQ questions that mapped to each | | | | | | | Total no. domains |
|---|---|------------------|-----------------------------|--------------------------------|-------------------------------|--|-----------------------------|-------------------|
| | Mind-body interactions in health and disease | Patient behavior | Physician role and behavior | Physician-patient interactions | cultural issues in healthcare | Social and health policy and economics | Health policy and economics | |
| Student perception of appropriate curricular time | N/A | 2 | 5 | 2 | 2 | 2 | N/A | 11 |
| Confidence with clinical knowledge and skills | N/A | N/A | 2 | 4 | 1 | 1 | N/A | 7 |
| Attitude toward medicine and profession | N/A | N/A | 4 | N/A | N/A | 2 | 2 | 6 |
| Total | N/A | 2 | 11 | 6 | 3 | 2 | 2 | 24 |

* Matrix subtype indicates the GQ question category that mapped to an SBS domain. See the text for an explanation of how and why the matrix was created. (N/A indicates not available.)

Table 2
Demographic Characteristics of Medical Students Who Completed the 2006 and 2007 AAMC Graduation Questionnaire, by Participating U.S. Medical School*

| Response rate; characteristic | Medical school and % students with each characteristic | | | | | | | | | |
|---|--|------|------|------|------|------|------|------|------|-----------------------------------|
| | A | B | C | D | E | F | G | H | I | % All schools with characteristic |
| Response rate | 83 | 100 | 66 | 83 | 78 | 83 | 60 | 48 | 65 | 74 |
| Gender: Male | 43.1 | 44.9 | 43.1 | 52.7 | 37.6 | 53.1 | 58.1 | 50.6 | 45.4 | 50.1 |
| Age: 26 or younger | 54.9 | 33.3 | 19.8 | 46.6 | 17.2 | 59.6 | 43.6 | 60.8 | 44.3 | 48.1 |
| Age: 27–29 | 33.6 | 47.8 | 50.5 | 34.2 | 50.5 | 30.2 | 40.0 | 30.4 | 35.8 | 36.7 |
| Age: 30–32 | 5.7 | 11.9 | 17.8 | 12.3 | 18.3 | 7.6 | 10.6 | 6.3 | 17.0 | 10.0 |
| Age: 33 or older | 5.7 | 6.9 | 11.9 | 6.8 | 14.0 | 2.7 | 5.9 | 2.5 | 2.8 | 5.3 |
| Race: Black | 2.4 | 8.8 | 2.9 | 12.3 | 3.2 | 2.7 | 5.8 | 6.3 | 2.8 | 6.1 |
| Race: White | 79.7 | 76.1 | 57.8 | 31.5 | 80.6 | 84.5 | 77.9 | 64.6 | 68.2 | 68.5 |
| Race: American Indian or Alaskan Native | 1.6 | 0.6 | — | — | 2.2 | 0.4 | — | 1.3 | 0.9 | 0.8 |
| Race: Asian | 8.9 | 12.6 | 28.4 | 34.2 | 14.0 | 7.5 | 9.3 | 26.6 | 18.7 | 16.9 |
| Race: Native Hawaiian or Pacific Islander | — | — | 1.0 | 2.1 | — | — | — | 1.3 | — | 0.7 |
| Race: Other | 11.4 | 3.1 | 16.7 | 24.7 | 2.2 | 5.8 | 9.3 | 5.1 | 13.1 | 10.4 |

* All but one school had 100–199 students, with one school at >200 students. Eight schools were urban and one school was rural. Three schools were private and six schools were public.

Table 3

Mean Scores, From Nine NIH-Grant Schools, for Institute of Medicine (IOM) Domains by Question Category of the 2006 and 2007 AAMC Graduation Questionnaire (GQ)

| Matrix subtype | | Mean (SD); range | Mean of all schools' scores |
|--|--|-----------------------|-----------------------------|
| IOM domain | GQ question category | | |
| Patient Behavior | Student Perception of Appropriate Curricular Time [*] | 75.4 (4.7); 68.2–81.9 | 77.3 |
| Physician Role and Behavior | Student Perception of Appropriate Curricular Time [*] | 81.5 (3.5); 75.6–86.4 | 84.4 |
| Physician Role and Behavior | Attitude Toward Medicine & Profession [†] | 60.8 (3.6); 56.5–66.0 | 61.4 |
| Physician Role and Behavior | Confidence with Clinical Knowledge and Skills [‡] | 93.4 (1.2); 91.5–94.9 | 93.5 |
| Physician–Patient Interactions | Student Perception of Appropriate Curricular Time [*] | 89.2 (5.1); 79.6–96.2 | 90.3 |
| Physician–Patient Interactions | Confidence with Clinical Knowledge and Skills [‡] | 86.5 (2.4); 83.2–90.3 | 88.3 |
| Social and Cultural Issues in Healthcare | Student Perception of Appropriate Curricular Time [*] | 73.4 (5.9); 65.2–85.0 | 74.9 |
| Social and Cultural Issues in Healthcare | Confidence with Clinical Knowledge and Skills [‡] | 89.8 (8.8); 70.9–98.4 | 92.2 |
| Health Policy and Economics | Attitude Toward Medicine & Profession [†] | 88.3 (4.1); 83.6–95.4 | 86.8 |

^{*} Student Perception of Appropriate Curricular Time = % appropriate ($\leq 2.8\%$ missing data).

[†] Attitude Toward Medicine and Profession = % strongly agree + % agree ($\leq 3.3\%$ missing data).

[‡] Confidence with Clinical Knowledge and Skills = % strongly agree + % agree ($\leq 1.1\%$ missing data).

Table 4

Sample Size Required Given a Range of Effect Sizes*

| Power | Tails | Alpha | Effect size | Standard deviation | M1 [†] | M2 [‡] | Difference between M2 and M1 | Sample size (n) |
|-------|-------|-------|-------------|--------------------|-----------------|-----------------|------------------------------|-----------------|
| 80% | 2 | .05 | 0.20 | 4.1 | 88.3 | 89.12 | 0.82 | 199 |
| 80% | 2 | .05 | 0.50 | 4.1 | 88.3 | 90.35 | 2.05 | 34 |
| 80% | 2 | .05 | 0.80 | 4.1 | 88.3 | 91.58 | 3.28 | 15 |
| 80% | 2 | .05 | 1.07 | 4.1 | 88.3 | 92.67 | 4.37 | 9 |

* The table's data indicate that a sample size of 34 schools would be needed to attain a medium effect size of 0.50 (50%). See the text for details.

[†] M1 is the mean for the Institute of Medicine domain of *Health Policy and Economics* and the Graduation Questionnaire question category of *Attitudes Toward Medicine and Profession*.

[‡] M2 is the mean given a sample size of 9, power of 80%, two tailed, alpha of .05, corresponding to an effect size of 1.07.