DOCUMENT RESUME

ED 410 263 TM 027 042

AUTHOR Scheuneman, Janice Dowd; And Others
TITLE An Evaluation of Gender Differences in Computer-Based Case Simulations.
PUB DATE Mar 97
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Case Method (Teaching Technique); Case Studies; Cognitive Processes; *Computer Simulation; Higher Education; Licensing Examinations (Professions); Medical Education; *Medical Students; Patients; *Performance Factors; *Sex Differences; *Test Items
IDENTIFIERS National Board of Medical Examiners

ABSTRACT

As part of the research leading to the implementation of computer-based case simulations (CCS) for the licensing examinations of the National Board of Medical Examiners, gender differences in performance were studied for one form consisting of 18 cases. A secondary purpose of the study was to note differences in style or approach that might differentiate the performance of men and women at a more detailed level than overall rating. In CCS, the examinee is presented with an introduction to the patient's signs and symptoms and then enters patient management plans into the computer. The simulated patient's condition changes in response to the action requested by the examinee. The sample in this study was 201 senior medical students. Gender identification was available for 118 men and 78 women. Performance on the total set of cases was similar for men and women, with the average of case means 4.55 for men and 4.51 for women. The two cases with the largest performance difference favoring women were obstetrics-gynecology cases, and an emergency surgery case had the largest difference favoring men. At the item level, results suggest that men tend to request more beneficial and inappropriate actions than women, although the effect was small. Overall, the performance differences on the CCS cases were very small, as expected. (Contains 6 tables and 10 references.) (SLD)

An Evaluation of Gender Differences in Computer-Based Case Simulations

Janice Dowd Scheuneman

Stephen G. Clyman

Yihua Van Fan

National Board of Medical Examiners

A Paper presented at the Meeting of the National Council on Measurement in Education

Chicago

March 1997

As part of the ongoing research leading to the implementation of National Board of Medical Examiners' (NBME)computer-based case simulations (CCS) as one component of the United States Medical Licensing Examinations (USMLE), gender differences in performance were investigated for one form consisting of 18 cases. An earlier study found gender differences with a set of four cases on which men performed significantly better overall (Orr, 1988). The present study uses both more recently developed cases covering a broader spectrum of the students' training and a different scoring model.

In general, a number of research studies have found that constructed response items show smaller differences between the performance of male and female examinees than multiple-choice items covering the same subject matter (Bell & Hay, 1987; Bolger & Kellaghan, 1990; Bridgeman & Lewis, 1994; Mazzeo, Schmitt, & Bleistein, 1993; and Murphy, 1982). Although little has been done investigating gender differences with more elaborate performance assessment measures, the expectation has been that these measures would similarly show reduced differences between men and women.

A secondary purpose of the study was to investigate differences in style or approach to problems that might differentiate the performance of men and women at a more detailed level than overall rating. Differences found might contribute to knowledge about gender differences either in response to the assessment instrument or in the practice of medicine generally that could have implications for other assessments for medical education.

Method

NBME's CCS are complex, unprompted, dynamic computer simulations of a patient-care environment. As a performance assessment instrument, they are intended to measure patient

management skills in a realistic environment with simulated time and naturally unfolding clinical situations.

In CCS, the examinee is initially presented with a brief introduction to the patient's signs and symptoms. No questions are asked of the examinee; he/she types requests for patient management into a blank screen, a process resembling the way orders are requested and processed in a hospital setting. The examinee can request a history, conduct a physical examination, write orders to perform diagnostic studies, initiate therapies, change the patient's location, or choose to perform no actions until a later time. The patient's condition changes in response to the actions requested by the examinee and the time course of any underlying disease. The examinee needs to decide when, where, and how to care for the patient through this evolving course. Each action requested by the examinee is recorded by the computer, including canceled or refused actions, the simulated time of the action, and the cost.

Data Source

The sample used in this study consisted of 201 senior medical students tested from 1991 to 1994 as part of a series of special studies. Gender identification was available for 196 of these students, 118 men and 78 women. A total of 18 cases from the areas of internal medicine, surgery, obstetrics/gynecology (ob/gyn), and pediatrics were administered in two separate sets. Each problem required an average of 20-25 minutes to complete.

CCS Scores

A committee of physicians provided holistic ratings of examinee performance on each case. The physicians were provided with transaction lists specifying the actions requested by the students in managing each case. Each transaction list was rated by two to six physicians on a

nine-point scale representing overall adequacy of the patient management. Definitions of the scale points were discussed and agreed upon by the raters prior to reading the transaction lists (Clauser et al, 1995).

In addition, analytic "items" were scored based on the examinee actions within each case. Item level differences were also considered in evaluating the overall performance differences between men and women. The items relate to the actions that might be requested by the examinee in managing the case and the appropriateness of those actions as identified by a key development committee consisting of physicians other than those assigning the holistic ratings. Actions that were appropriate in the management of the patient were designated "benefit" actions. Actions that were not beneficial were further divided into actions that are inappropriate but harmless, actions that present some risk to the patient, and flags that represent serious errors in management that could severely jeopardize the patient. Neutral actions, those which were neither beneficial nor inappropriate, were also identified but not evaluated as part of this study.

Results

The performance on the total set of cases was very similar for the men and women in this data set. The average of the case means for men was 4.55 and for women was 4.51. Standard deviation of the case means was about 0.6 for both groups. The median of the standard deviations of ratings for the 18 individual cases was 1.56 for men and 1.52 for women.

To provide some context for these results, we also obtained multiple-choice scores from Part III of the NBME examinations. (Part III was the last of the three parts making up the predecessor to the USMLE. It was given for the last time in May 1994). The Part III examinations were selected from among the NBME examinations because the content coverage was most similar to

that measured by the CCS. Scores were found in the NBME research data base for 96 of the men and 62 of the women in the CCS sample. Mean scores were again very similar for men and women (459 for men and 455 for women), but standard deviations were much larger for men than women, 110 and 86 respectively. This difference in standard deviations is not observed in Part III data generally, although overall mean differences are similar to those observed in this study (Dillon et al, 1995). Summary data for CCS ratings were recomputed for only that part of the sample for whom Part III scores were available. Results were essentially the same as the results for the full sample that are given above, indicating that the sample with Part III scores was a representative subset of the whole.

Mean ratings for each case were also compared for men and women. Although the majority of cases showed small mean differences, the range of differences extended from -.88 (men performed better) to .48 (women performed better). These differences are about 0.5 and 0.3 standard deviation units respectively. The distribution of rating differences is shown in Table 1.

Performance by Area of Medicine

Medical disciplines, such as internal medicine, surgery, ob/gyn, and pediatrics, represent differences in the age and gender of the patient and often the degree and type of intervention required. They are used here as convenient proxies for these variations in cases. Previous research with the NBME examinations and the USMLE has suggested that women tend to perform better in ob/gyn, while men perform better in internal medicine and surgery (Case, Becker, & Swanson, 1993). Patterns of differences by area of medicine were also of interest for the CCS cases. Because of the low generalizability that has been reported for individual performance assessments, the degree to which the different cases from the same are of medicine

Table 1
Distribution of Difference between Means Scores of
Men and Women on CCS Cases

Difference Between Means	Number of Cases
>.40	1
.31 to .40	1
.21 to .30	1
.11 to .20	2
0.0 to .10	2
-0.0 to -.10	5
-.11 to -.20	3
-.21 to -.30	2
-.31 to -.30	0
<-.40	1

form coherent sets was first examined. Correlations among the 18 problems and with the Part III scores were therefore computed. Table 2 provides the median correlation among cases from the same area of medicine and between areas of medicine and between the cases and the Part III scores.

Cases in medicine, pediatrics, and ob/gyn show some internal consistency, although they are not clearly different from each other. The median of correlations among cases of the same type are only slightly higher than those between cases areas, but differences are in the right direction. Surgery cases, however, do not appear to be correlated among themselves as highly as with cases from the other areas, suggesting that interpretations based on the surgery content may not be

Table 2
Median Correlations between Cases by Area of Medicine
and between Cases and Part III Scores[1]

	Int. Medicine	Ob/Gyn	Pediatrics	Surgery
Int. Medicine	.23 (15)	.20 (30)	.22 (24)	.21 (18)
Ob/Gyn		.22 (10)	.15 (20)	.17 (15)
Pediatrics			.25 (6)	.19 (12)
Surgery				.09 (3)
Part III	.27 (6)	.16 (5)	.28 (4)	.13 (3)

warranted. Although only three surgery cases are included in the set of 18, this is only one less than for pediatrics, which shows results similar to those for medicine with its six cases.

Looking at the cases with the largest performance differences between men and women, two of the three cases with largest differences favoring women were ob/gyn cases while the case with the largest difference favoring men was an emergency surgery case. In fact, women performed better than men to some degree on four of the five ob/gyn cases. Men performed better to some degree on all three of the surgery cases and on five of the six internal medicine cases. Little difference in performance between men and women was seen on the pediatric cases. Means and differences between means for different cases are shown in Table 3.

[1] Number of correlations in the cell is given in parentheses.

Table 3
Performance of Men and Women by Case Area

	Medicine	Surgery	Pediatrics	Ob/Gyn	Overall
N cases	6	3	4	5	18
Mean Men	4.58	4.70	4.21	4.69	4.55
Mean Women	4.40	4.53	4.19	4.90	4.51
sd--pooled	1.16	.88	.96	.88	.79
Difference (sd units)	-.16	-.18	-.02	.23	-.05
Range of Case Differences	-.88 to .26	-.29 to -.02	-.15 to .08	-.05 to .48	-.88 to .48

Item Level Performance

The performance of men and women was also compared with regard to beneficial, inappropriate, risk, and flag actions requested. Overall, men requested more beneficial and inappropriate actions, although the differences were quite small. Results are shown in Table 4. On a case by case basis, however, men requested more benefit and inappropriate actions on 15 of the 18 cases and more risk actions on 14 of the 18 cases, suggesting that some tendency exists for men to request slightly more actions than women.

Also shown in Table 4 are correlations of the different numbers of actions requested with the ratings. The negative correlations should probably be interpreted to mean that cases that elicit many requests for benefit actions may be more difficult to manage appropriately. The association of more risk actions with poorer performance seems appropriate, although previous

Table 4
Number of Actions Requested and
Correlations with Ratings by Type of Item
Separately for Men and Women

	Men	Women
Mean Numbers of Actions		
Beneficial Actions	10.60	10.30
Inappropriate Actions	.89	.78
Risk Actions	.24	.22
Flag Actions	.07	.08
Total Actions	11.81	11.37
Correlation with Rating		
Benefit Actions	-.18	-.39
Inappropriate Actions	-.09	-.04
Risk Actions	-.81	-.67

research has shown that cases that have many possibilities for risk actions also tend to be more difficult (Scheuneman, Fan, & Clyman, in press). The difference in the magnitude of the correlations for men and women is of some interest. The association of difficulty and number of benefit actions requested is higher for women, while the association with difficulty and number of risk actions requested is higher for men.

The correlations between action type and the difference between ratings for men and women were also computed. These correlations suggest further that cases with many benefits tend to favor men while cases with many risks tend to favor women. This raises the question of whether

Table 5
Number of Actions Requested by Type of Case
Separately for Men and Women

		Benefit	Inappropriate	Risk
Medicine	Men	9.71	1.16	.23
	Women	9.30	1.12	.18
Surgery	Men	14.43	.88	.04
	Women	13.72	.69	.05
Pediatrics	Men	12.13	1.13	.36
	Women	11.88	.94	.31
Ob/Gyn	Men	8.14	.38	.28
	Women	8.20	.29	.28

this result occurs because kinds of actions are associated with different case types. The number of actions requested by area of medicine is shown in Table 5. This shows that the association with differences between areas of medicine in types of actions requested does not correspond with the gender differences by areas of medicine. Surgery and pediatrics cases average more benefit actions and surgery cases average fewer risk actions. Fewer risks for surgery may seem surprising as surgery seems to be a riskier enterprise in general than pediatrics or ob/gyn. Remember, however, that actions that are potentially harmful to the patient but are appropriate for patient management are considered benefits, a situation that may more often apply to surgery cases. Notice also that, although women generally requested fewer benefit actions than men, they requested more benefit actions for ob/gyn cases.

Table 6
Mean Number of Items Requested by Men and Women
for the Six Cases with largest differences[2]

Content	Mean Difference		Items		
			Benefit	Inappropriate	Risk
Pulmonary	-.88	Men	13.26	.47	.17
		Women	11.76	.51	.21
Trauma	-.29	Men	16.44	.96	.06
		Women	15.09	.87	.05
Gastrointestinal	-.24	Men	6.90	.93	.10
		Women	6.69	.92	.10
Metabolic	.26	Men	7.08	2.19	.36
		Women	6.97	1.82	.14
Complication of Pregnancy	.36	Men	6.88	.07	.45
		Women	6.85	.12	.64
Pregnancy	.48	Men	9.75	.20	.08
		Women	9.63	.06	.13

Examination of Cases with Large Gender Differences

The six cases with the largest performance differences between men and women, three favoring men and three favoring women, were examined for possible patterns of differences. Of the three cases favoring women, two were ob/gyn cases and one was a medicine case with a metabolic condition more often seen in women. The cases with the largest difference favoring men were an emergency trauma case, classified as surgery and two internal medicine cases, one with a pulmonary problem, the other gastrointestinal.

[2] Negative differences indicates that men performed better; positive differences that women performed better.

The number of actions requested for these six cases are shown in Table 6. For the two cases with the largest differences favoring , men, men requested more benefit actions than women. At the individual item level, several of the benefit items were requested by a higher proportion of men than women. Some of those benefits were appropriate actions that carried a risk for harm to the patient. For the metabolic case, men requested somewhat more inappropriate and risk actions. Otherwise, the difference in actions requested were small or unexpected. For example, women requested slightly more risk actions on the pregnancy complication case. Apparently the differences in management that resulted in the mean differences in ratings were not always reflected in the individual items requested. This could be due to combinations of actions or timing of actions that are not reflected in the simple counts of requests made.

Discussion

Overall, the performance differences on the CCS cases were very small as expected. Performance differed somewhat, however, according to the area of medicine, with women generally performing better on ob/gyn cases and men on medicine cases. Men also performed better on the set of cases classified as surgery cases, although the low intercorrelations among ratings on these cases suggest that this may not be a generalizable result.

At the item level, results suggest that men tend to request more actions than women, although the effect was small. Some of the results suggest that men may be more likely to request actions that pose a risk to the patient. If these actions are appropriate, the result may be more effective management; if not appropriate, the management may be less effective. Since potentially harmful actions may be either appropriate or inappropriate, this can not be readily

determined from the data. Benefit actions could perhaps be rated as potentially harmful or harmless to determine if this hypothesis may have merit.

The importance of the content blueprint for assembling sets of cases for future CCS forms is confirmed by the results showing some gender differences according to area of medicine. The balance on the set of cases used in this study was good and the content blueprint should help retain an equitable balance in the future. Some deliberation might be appropriate, however, concerning cases that might show fairly large differences such as the pulmonary case in this set. Should screening for such cases be instituted? If differences are found, what action should be taken?

Overall, the evaluation of the CCS for purposes of inclusion in the USMLE must be considered positive. Any gender differences in management of the cases do not in general appear to affect performance ratings. The CCS appears to be a fair and equitable measure for use in this important examination for medical licensure.

References

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. British Journal of Educational Psychology, 57, 212-220.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27, 165-174.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice score with grades in college courses. Journal of Educational Measurement, 31, 37-50.

Case, S. M., Becker, D. F., & Swanson, D. B. (1993). Performance of men and women on NBME Part I and Part II: The more thing change... Academic Medicine, 68 (October Supplement), S25-S27.

Clauser, B. E., Subhiyah, R. G., Nungester, R. J., Ripley, D. R., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. Journal of Educational Measurement, 32, 397-415.

Dillon, G. F., Henzel, T. R., LaDuca A., Walsh, W. P. (1995). The influence of type of residency training and gender on an examination for medical licensure. Paper presented at the Annual Meeting of the American Education Research Association, April, San Francisco.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). Sex-related differences on constructed response and multiple-choice sections of Advanced Placement Examinations (Report No. 92-7). New York: College Entrance Examination Board.

Murphy, R. J. L. (1982). Sex differences in objective test performance. British Journal of Educational Psychology, 52, 213-219.

Orr, N. (1988). Non-cognitive correlates of performance on conventional and interactive components of a computerized examination. Research in Medical Education -Proceedings of the Twenty-Seventh Annual Conference, 284-289. Washington, DC: Association of American Medical Colleges.

Scheuneman, J. D., Fan, Y. V., & Clyman, S. G. (In press). An Investigation of the Difficulty of Computer-Based Case Simulations. Medical Education.

NCME 1997

TM027042

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Evaluation of Gender Differences in Computer-Based Case Simulations	
Author(s): Janice Scheuneman, Stephen Clyman, Yihua Van Fan	
Corporate Source: National Board of Medical Examiners	Publication Date: 3/97

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Janice Scheuneman	Position: Senior Evaluation Officer
Printed Name: Janice D. Scheuneman, PhD	Organization: National Board of Medical Examiners
Address: 3750 Market Street Philadelphia, Pa. 19104	Telephone Number: (215) 590-9669
	Date: April 8, 1997

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 24, 1997

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our process of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: NCME 1997/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an NCME chair or discussant, please save this form for future use.