International Journal of Technology Assessment in Health Care, 23:1 (2007), 108–115.

An experimental study of the influence of individual participant characteristics on formal consensus development

James Carpenter, Andrew Hutchings
London School of Hygiene & Tropical Medicine

Rosalind Raine
University College London

Colin Sanderson
London School of Hygiene & Tropical Medicine

Objectives: The aim of this study was to examine the influence of participants' characteristics on the results produced by formal consensus methods.
Methods: The approach was an experimental study of 346 participants in 20 groups rating the appropriateness of four mental health interventions for the treatment of chronic fatigue syndrome, irritable bowel syndrome, and chronic back pain. There were four factors in the design: systematic literature review provided or not, decisions made under realistic or "ideal" resource assumptions, clinically mixed (general practitioners and mental health professionals) or homogenous group (general practitioners only), convened or mail-only group. A group's rating was defined as the median of participants' ratings. The influence of participants' characteristics (age, sex, and specialty) was examined using multilevel models.
Results: The largest differences were between the GPs and mental health professionals, both in their initial ratings of the different interventions, and in how much they altered their ratings between rounds. There were smaller but statistically significant ($p < .05$) differences between specialty and age groups in initial ratings for the treatment (by whatever means) of different conditions, and for certain conditions women increased their ratings more than men. Women rated intervention more favorably when assuming "ideal" rather than realistic levels of resources, but men did not.
Conclusions: Our findings support the practice of treating professional specialty as an important determinant of the results in consensus panels.

Keywords: Consensus methods, Delphi technique, Professional role, Sex, Age groups

Financial support for this study was provided entirely by a grant (MRC Clinician Scientist Fellowship for R. Raine) from the Medical Research Council. The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. This study is part of a research program examining the methodology of group decision making for clinical guideline development. This research program is overseen by a steering committee comprising one of the authors (C.S.), Nick Black, Tom Sensky, Andy Haines, and Simon Carter. We thank the general practitioners and mental health specialists who participated in the study, and Kirsten Larkin for administrative support. The study received ethical approval from the London School of Hygiene & Tropical Medicine Ethics Committee.

Treatment recommendations in clinical practice guidelines are commonly based on formal consensus methods, based in turn on a combination of the best available scientific evidence and expert judgment (2). Typically consensus development methods involve two or more stages or "rounds," with clinical experts making individual judgments in private before, and then again after, exposure to the views of others.

In one of these methods, the nominal group technique (NGT), participants convene for a facilitated discussion of their views (14). An alternative is the mail-only Delphi survey, in which the summarized results of successive rounds are sent back to participants so that they can revise their own opinions in the light of others' (14). The RAND/UCLA appropriateness method combines elements of both the NGT and Delphi, with an expert panel that rates the appropriateness (1) or necessity (9) of interventions before and after a facilitated meeting. These methods differ from National Institutes of Health consensus development conferences in four ways: decisions are elicited in private, there is formal feedback of group views, interaction is structured, and explicit methods are used for the aggregation of views (14).

Many studies have examined the impact of clinical specialty and nationality on the views produced in formal consensus development (6). These studies show that providers of interventions tend to rate those interventions more favorably than nonproviders. No consistent patterns were observed by nationality. A methodological weakness with much of this research has been that the many ratings involved are often treated as independent observations in the analysis. This ignores any correlation in an individual participant's ratings of multiple scenarios (participant level effects) and any correlation in the ratings of participants within the same group (group level effects). Failure to take level-specific effects into account appropriately by treating the participant or group as the unit of analysis, for example, by treating participants or groups as random effects, generally leads to underestimated standard errors.

The only studies of the possible effects of participant characteristics on how ratings *alter* between rounds have examined the specialty of participants. Most studies found that differences between "providers" and other clinicians tended to remain similar or get smaller (4;10;15;20), although there have been some exceptions, depending on the intervention being rated (15), and in one study, the views of a mixed and a surgeon-only group diverged (19).

In other areas of clinical decision making, studies have shown that physician attributes such as age, gender, ethnicity, and experience affect judgments (3;11;12). However, there is very little evidence about the effects of factors of this kind on the results of consensus development processes. In this study, we examine the relationships between participant characteristics, such as age and gender, on the ratings and changes in ratings between rounds in consensus development processes, and revisit the effect of profession.

METHODS

The data came from a research program conducted in England, which involved sixteen convened groups and four mail-only groups in a factorial design (Figure 1). The convened groups differed with respect to three design factors: group composition (general practitioners [GPs] only or mixed GPs and mental health professionals [MHPs]), provision of a literature review (provided or not), and decision context ("realistic" assumptions about levels of resources available in the UK National Health Service versus "ideal" levels of resources). The mail-only groups differed with respect to group composition only; all were provided with a literature review and made decisions assuming a "realistic" resource context. Details of the group-level design factors and their impact on ratings at the group level have been reported elsewhere (7;8;17).

We selected three conditions (chronic back pain [CBP], irritable bowel syndrome [IBS], and chronic fatigue syndrome [CFS]) for study. These conditions were chosen because they were (i) common somatic conditions for which GPs had indicated they would welcome guidance, but no

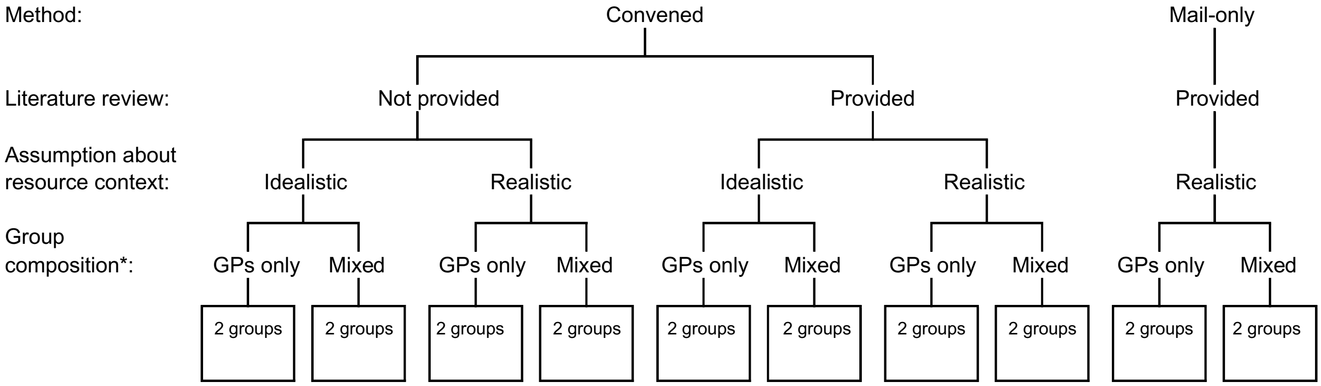

Figure 1. Study design. The asterisk indicates general practitioners (GPs) or mixed (GPs and mental health professionals).

national guidelines existed; (ii) cared for by at least two groups of clinicians (GPs and MHPs); and (iii) commonly treated in the United Kingdom in ways that were inconsistent with the research evidence (for example, antidepressants were used for CFS despite the lack of research evidence for their effectiveness).

We conducted a systematic review of the effectiveness of mental health interventions in primary care for patients with these conditions (16). Four relevant interventions were identified: behavioral therapy, cognitive behavioral therapy (CBT), brief psychodynamic interpersonal therapy, and antidepressants.

A questionnaire covering sixty-four clinical scenarios was developed. This elicited the views of the participating GPs and MHPs about appropriate use of the four interventions for improving physical functioning in working-age adults for the three conditions studied, in the presence or absence of four cues identified as relevant by clinicians (17). The cues were coexistent depressive symptoms (in all conditions), back pain induced insomnia (in patients with CBP), a financial motivation to return to work (in patients with CFS), and patient belief that their condition had an organic cause (in all conditions). For example, one scenario was "When a patient suffering from irritable bowel syndrome demonstrates symptoms of depression, cognitive behavioural therapy is a good treatment option to improve physical functioning." Participants were asked to rate their level of agreement on Likert scales from 1 (strongly disagree) to 9 (strongly agree). Participants were provided with detailed descriptions of the conditions and interventions (17).

The participants were drawn from national random samples of GPs from the Department of Health GP database for England ($N = 27{,}723$), and of MHPs from the Royal College of Psychiatrists liaison section database and the British Association of Behavioural and Cognitive Psychotherapists database (total $N = 720$).

The target number of eleven participants in each convened group was based on the literature (18). We invited 2,680 GPs and 310 MHPs to take part. Invitations were sent 2 months before each meeting, and the first 14 responders (stratified for the mixed groups) were recruited to allow for attrition. They were sent a questionnaire and completed their first round of ratings by mail. For the second round, each group convened for a facilitated meeting, which followed a written protocol. At the meeting, each participant was given a new copy of the questionnaire that included a reminder of their own initial ratings and the distribution of ratings for the whole group. Each scenario was discussed in turn, and reasons for any differences were explored. The participants then privately re-rated each scenario. We also invited another 2,000 GPs and 410 MHPs to participate in the mail-only groups. All those who agreed to participate and returned completed first-round ratings were randomly allocated to one of four mail-only groups (forty-six per group). For the second round, they were sent a new copy of the questionnaire with a reminder of their own initial ratings, and the distribution of ratings for the whole group.

Analysis

The focus was on exploring the factors that affect (i) the distances of participants' ratings from their group's median rating at round 1, and (ii) participant's changes in response between round 1 and 2. Specifically, we explored the effect of interactions between participant-specific variables and design variables using a multilevel model for each participant's answers to the sixty-four scenarios at each round (5).

Three participant and six design variables were used in the analysis. The participant-specific variables (reference categories in parentheses) were sex (female), age (centered at 45 years), and job (mental health professional). The group-specific design variables (reference categories) were type of process (convened group), group composition (GPs only), literature review (provided), and resource context ("realistic"). The scenario-specific design variables (reference categories) were the intervention (CBT) and the condition (CFS).

A group's rating for a question was defined as the median of its members' ratings. For analysis 1, the factors affecting the distance of individual ratings from the group median in round 1, models were fitted by maximum likelihood using the *MLwiN* software (13). As the study design was hierarchical (scenario factors within participant factors within group factors), the statistical analysis began by investigating components of variance at each level of the hierarchy. Non-zero variance was only found at the scenario and participant level. Accordingly, all models have variance components at these two levels.

Because the underlying data come from bounded scales, distance from a group's median rating is potentially affected by the position of this rating because of floor and ceiling effects. All models were adjusted for group median and its square (providing a greater adjustment close to the floor/ceiling). Next, fixed effects for all participant and design variables were added to the model. Each possible interaction between the three participant and six design variables was investigated, and its statistical significance was evaluated using Wald's tests. The final model was derived by including all participant and design variables, and adding interactions one at a time, starting with the most statistically significant, until (i) no more interactions were significant at the 5 percent level and (ii) all the interactions in the model were significant at the 5 percent level.

In analysis 2, the dependent variable was the difference between individual ratings in rounds 1 and 2. A positive value indicated increased support for an intervention. The focus was on investigating the effect of participant characteristics and interactions between participant and design variables on this change. However, change in opinion is likely to be affected by the distance from, and level of, the group rating at round 1, so these were adjusted for in all analyses.

Non-zero variance was found at the group, participant, and scenario level, so components of variance for all three were included in all analyses. Interactions were investigated using the same model-building strategy as Analysis 1. The extent to which participants' opinions carried over from round 1 to round 2 was examined by comparing the participant-specific residuals at the scenario and participant levels. Model assumptions were checked by examining normal probability plots of residuals at all levels, and plots of scenario level residuals against fitted values.

RESULTS

Participants

Participation rates at round 1 were 6.8 percent (5.8 percent for GPs, 15.5 percent for MHPs) in the convened groups and 8.8 percent (7.2 percent for GPs, 16.8 percent for MHPs) in the mail-only groups (8). Round 2 ratings were completed by 90.6 percent of the round 1 participants. There were 177 participants (135 GPs and 42 MHPs) in the 16 convened groups and 169 participants (128 GPs and 41 MHPs) in the four mail-only groups at round 2. All analyses were based on data from participants who completed ratings for both rounds.

Mean age was 46.1 years (standard deviation 4.6) and was similar for GPs and MHPs (45.8 and 46.2, respectively, $p = .67$). There were 125 females, with a lower percentage among the GPs than the MHPs (30.4 percent and 54.2 percent, $p < .001$). Mean age was similar for women and men (45.3 and 46.6, respectively, $p = .20$). The mean number of years of relevant experience was 17.0 (standard deviation 7.78) and this did not differ significantly between GPs and MHPs (16.9 and 17.6, respectively, $p = .48$).

Factors Influencing Individual Participant's Variation around Their Group Rating for Each Scenario at Round 1

The coefficients of the model for distance from the group rating at round 1 are shown in Table 1. Examination of the residual plots at the scenario and individual level indicated good agreement with the assumptions of normality and constant variance.

It was found that all significant (p < .05) main effects were included in interactions. Therefore, the concern here is with interactions between effects: to what extent were the ratings for specific conditions and interventions different for different types of participant?

Ratings of the different interventions depended on the participant's profession (Table 1). MHPs were generally much more in favor of cognitive behavioral therapy and behavioral therapy than GPs ($p < .001$, both comparisons). GPs were generally more in favor of brief psychodynamic

Table 1. Estimated Coefficients for Model Describing Distance from Consensus at Round 1

	Estimate (*SE*)	*p* value[a]		Estimate (*SE*)	*p* value[a]
Fixed coefficients: main effects			**Interactions of subject specific and design effects**		
Intercept[b]	1.021 (.149)	<.001	*Job and Therapy*		
Group median rating, centered at 5	−.297 (.011)	<.001	GP and behavioral therapy	.568 (.073)	Joint test:
Square of group median rating	.022 (.006)	<.001	GP and BPIT	1.498 (.073)	<.001
Method (reference convened groups)	−.023 (.122)	.850	GP and antidepressants	1.509 (.073)	
Group mix (reference: MHPs and GPs)	.176 (.104)	.091	*Age and Condition*		
Literature review (reference: provided)	.046 (.122)	.706	Age and IBS	.014 (.003)	Joint test:
Effects included in interactions			Age and CBP	.008 (.003)	<.001
Sex (reference: female)	−.033 (.108)	.760	*Sex and Context*		
Age, centered at 45 years	−.006 (.005)	.230	Male and Ideal context	−.436 (.202)	.031
Job (reference: MHP)	−1.208 (.135)	<.001	*Job and Condition*		
Context (reference: realistic)	.321 (.174)	.065	GP and IBS	−.180 (.066)	Joint test:
Behavioral therapy (reference: CBT)	−.501 (.063)	<.001	GP and CBP	−.016 (.059)	.014
BPIT (reference: CBT)	−1.587 (.071)	<.001	**Variance components**		
Antidepressants (reference: CBT)	−1.389 (.068)	<.001	Between participants	.603 (.049)	<.001
Patient has IBS (reference: CFS)	.122 (.058)	.035	Between scenarios	2.655 (.025)	<.001
Patient has CBP (reference: CFS)	−.015 (.052)	.773			

[a] All *p* values were calculated from Wald's tests, except for variance components, where the likelihood ratio test was used with reference distribution of $.5(\chi_0^2 + \chi_1^2)$.

[b] The intercept represents the mean distance from the group median rating at the reference values of the covariates (45 years old, female mental health professional, in a mixed, convened group provided with a literature review, assuming a "realistic" resource context, rating cognitive behavioral therapy [CBT] in chronic fatigue syndrome [CFS], and with a group rating of 5 on the Likert scale). The intercept or mean distance from the group rating for all reference categories was approximately 1, partly because the distance from the median in the reference categories is positively skewed; opinions in favor of treatment tended to be stronger than those against. As expected, however, the distribution of variation about the median was related to the median value, the average distance from the median moving from above to below the median as the median value increased above 5.

MHP, mental health professional; GP, general practitioner; BPIT, brief psychodynamic interpersonal therapy; IBS, irritable bowel syndrome.

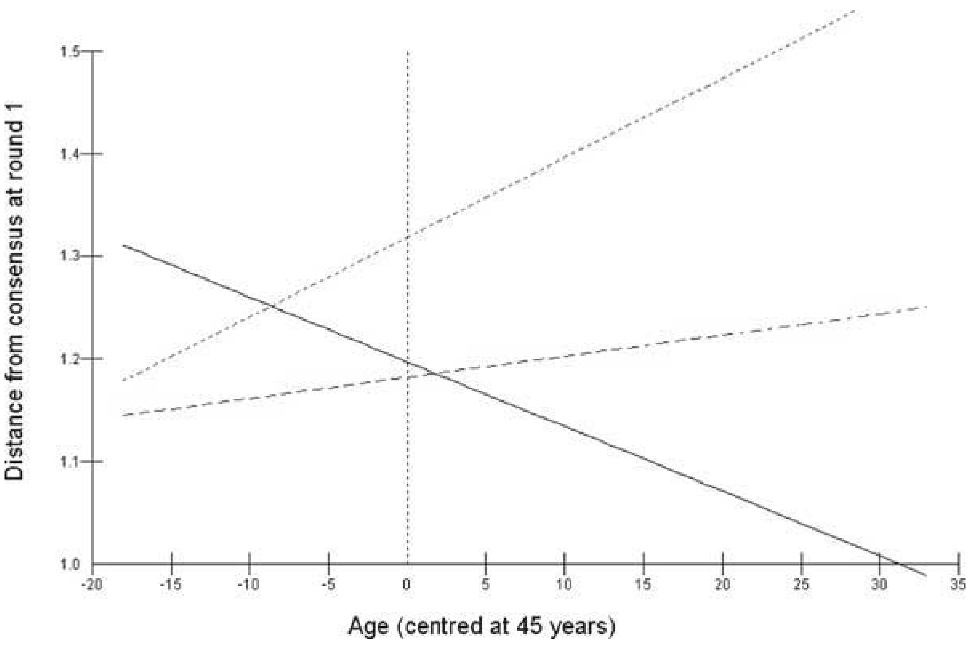

Figure 2. Effect of age on distance from consensus at round 1. Solid line, chronic fatigue syndrome; short dashed line, irritable bowel syndrome; long dashed line, chronic back pain.

therapy and antidepressants than MHPs ($p = .031$ and .026, respectively), but they were divided about the value of these (both p values against null hypothesis of being equally in favor/against > .05), with particular doubts about their value for irritable bowel syndrome ($p = .018$).

Responses to resource context depended on the participant's gender. Men and women provided similar ratings in a "realistic" context, but only women rated interventions more favorably in an "ideal" context ($p = .031$).

Ratings for treating different conditions depended on the participant's age (Figure 2). Increasing age was associated with a tendency to rate interventions for irritable bowel syndrome more favorably and interventions for chronic fatigue syndrome less so, with a significant ($p < .05$) difference between ratings for the two conditions in those 45 years of age and older.

Factors Influencing the Change in Individual Participant's Scores for Each Scenario between Rounds

The coefficients of the model for change in response between round 1 and 2 are shown in Table 2. Examination of residuals indicated some evidence of underdispersion relative to the normal distribution, particularly at the scenario level. However, given the large data set, the distribution of the estimators is likely to be close to normal, so this effect is unlikely to affect inferences materially. Significant components of variance were found between scenarios, individuals, and groups ($p = .04$ for the latter).

- In terms of main effects, the average change in the reference group was approximately .4 points in favor of treatment;
- As expected, the distance from the round 1 median affected this finding. In broad terms, for every point above (below) the median an individual was at round 1, their rating in round 2 was .3 less (more) in favor of treatment, suggesting a tendency for the spread of individual ratings to reduce;
- The group median in round 1 did have a slight effect on the mean change between rounds. For each point the median was above (below) 5, the average rating in round 2 was .06 points lower (higher), suggesting a slight tendency for group ratings to move toward the middle of the range;
- The mean change in favor of treatment was .15 points smaller in the mail-only groups than the convened groups ($p = .024$), and with no literature review, the change in favor of treatment was .16 points smaller ($p = .007$).

Changes between rounds in ratings of the different interventions depended on the participant's profession. For cognitive behavioral therapy, MHPs changed their ratings on average .12 points more in favor of treatment than GPs ($p = .04$). For behavioral therapy, MHPs and GPs both became on average .13 more in favor of treatment; for antidepressants, they both became on average .20 points more in favor. For brief psychodynamic therapy, GPs did not change their mean

Table 2. Estimated Coefficients for Model 2, Describing the Difference between Round 2 Score and Round 1 Score

	Estimate (*SE*)	*p* value[a]		Estimate (*SE*)	*p* value[a]
Fixed coefficients: main effects			**Interactions of subject-specific and design effects**		
Intercept[b]	.405 (.084)	<.001	*Job and Therapy*		
Distance from group median rating at round 1	−.303 (.004)	<.001	GP and behavioral therapy	.132 (.045)	Joint test:
Square of group median rating	.011 (.001)	<.001	GP and BPIT	.319 (.046)	<.001
Method (reference: convened groups)	−.152 (.068)	.024	GP and Antidepressants	.172 (.046)	
Group mix (reference: MHPs and GPs)	.046 (.054)	.398	*Sex and Condition*		
Literature review (reference: provided)	−.163 (.060)	.007	Male and IBS	−.131 (.037)	Joint test:
Age, centered at 45 years	−.002 (.002)	.334	Male and CBP	−.077 (.033)	.001
Effects included in interactions			**Variance components**		
Sex (reference: female)	−.007 (.042)	.869	Between groups	.005 (.004)	.043
Job (reference: MHP)	−.119 (.057)	.037	Between participants	.088 (.008)	<.001
Behavioral therapy (reference: CBT)	−.281 (.040)	<.001	Between scenarios	1.033 (.010)	<.001
BPIT (reference: CBT)	−.613 (.045)	<.001			
Antidepressants (reference: CBT)	−.224 (.043)	<.001			
Patient has IBS (reference: CFS)	.150 (.029)	<.001			
Patient has CBP (reference: CFS)	.105 (.026)	<.001			

[a] All *p* values were calculated from Wald's tests, except for variance components, where the likelihood ratio test was used with reference distribution of $.5(\chi_0^2 + \chi_1^2)$.

[b] The intercept represents the mean change in response from round 1 to round 2 (positive values indicate greater agreement with the intervention) at the reference values of the covariates (45 years old, female mental health professional, in a mixed, convened group provided with a literature review, assuming a 'realistic' resource context, rating cognitive behavioral therapy [CBT] in chronic fatigue syndrome [CFS], and with a group rating of 5 on the Likert scale in round 1).

MHP, mental health professional; GP, general practitioner; BPIT, brief psychodynamic interpersonal therapy; IBS, irritable bowel syndrome.

ratings, but MHPs became on average .21 points less in favor ($p <$ .001).

Changes between rounds in ratings of treatments for specific conditions depended on gender. For the baseline category, chronic fatigue syndrome, the men's and the women's ratings both moved on average approximately .40 points toward greater agreement with treatment. However, the women's ratings for chronic back pain and irritable bowel syndrome shifted more than this (chronic back pain: .51 versus .43, $p = .04$; irritable bowel syndrome: .56 versus .42, $p = .002$), and the shifts in men's ratings were similar for all three conditions.

The regression coefficient for distance from consensus at round 1 was −.3, consistent with a correlation between round 1 and 2 scores of approximately .7 after adjustment for other variables. If round 2 score were independent of round 1 score, this coefficient would be −1. (Note: let Y_2 be the response at round 2 and Y_1 the response at round 1. Analysis 2 regresses the differences between these on distance from consensus at round 1 [denoted c_1] and other covariates. Algebraically, $(Y_2 - Y_1) = \beta_0 + \beta_1(Y_1 - c_1) + \cdots$ so if the response at round 2 is independent of the response at round 1, Y_1 will cancel out of this equation; for this to happen $\beta_1 = -1$.) The implication is that much of the variation in round 2 score can be predicted by round 1 score, as can be seen from the variance components for scenarios and participants in analysis 2, which are much smaller than in analysis 1.

The correlation between the person-specific residuals at rounds 1 and 2 was .36 ($p < .001$). After adjustment for other factors (including distance from consensus at round 1), participants who were above consensus at round 1 tended to increase their ratings at round 2 more than those who were not, although this result is generated by relatively few participants. However, this effect is counterbalanced by the coefficient in model 2 for "distance from consensus at round 1" being −.3. Taken together, this means that those who were further above/below consensus at round 1 tended to reduce/increase their ratings at round 2 by less than the fixed coefficient of −.3 per unit in the round 2 model suggests. This finding is consistent with an overall tendency for round 2 ratings to move closer to consensus, but for this tendency to be proportionately smaller in those holding more extreme views at round 1. The correlation between the scenario-level residuals in the two models is −.0083 ($p = .26$), consistent with no overall residual scenario effect across rounds.

DISCUSSION

Interpretation of Results

We investigated how individual characteristics of participants interact with design variables to affect (i) distances from group consensus in round 1, and (ii) changes in opinion between the first and second rounds. In terms of design or group-level factors, the consensus method, provision of literature review, or group composition had no significant main effects on median ratings. For group composition, this can be interpreted as showing that the GPs in the mixed groups provided similar ratings to the GPs in the GP-only groups. Providing a literature review and using the Delphi consensus method both reduced the changes in ratings between rounds 1 and 2, but only by approximately .15 points on the nine-point scale.

In terms of individual-level factors, there was clear evidence of initial differences of opinion between GPs and MHPs about treatment options. In general, MHPs were more in favor of cognitive behavioral therapy and behavioral therapy than GPs, and less in favor of brief psychodynamic therapy and antidepressants. These results are likely to reflect the professional experience of participants with these conditions. Qualitative analysis of transcribed audiotapes of the convened groups in our study suggests that although GPs and MHPs tended not to have extensive experience of brief psychodynamic therapy, some GPs said they applied some of its principles in their everyday practice (unpublished data, 2004).

Women were more responsive than men to assumptions about the levels of resources available for treatment. The reason for this finding is unclear and requires qualitative investigation if it is to be used to inform the composition of formal consensus development groups in the future.

Age (or seniority) affected ratings for different conditions. Older panelists were less in favor of using mental health interventions in the management of chronic fatigue syndrome than for irritable bowel syndrome and chronic back pain. Qualitative data suggested that some panelists were skeptical about the value of any intervention for chronic fatigue syndrome. This was usually linked to doubts as to whether the condition existed at all and is consistent with these beliefs being more prevalent among older professionals (17).

Overall, the strongest predictor of how much an individual's answer to a question changed between rounds was distance from consensus at round 1. In addition, the amount of change in ratings for different treatments varied between professional groups, and women increased their ratings slightly more than men for interventions for chronic back pain and irritable bowel syndrome.

Methodological Issues

One major strength of this study is its size. Data from 346 participants in 20 groups allowed us to estimate differences with greater precision than previous studies, while taking account of the hierarchical nature of the data.

Our panel members were drawn from populations of practitioners who work with patients with these conditions as part of their daily practice. However, they differed from many guideline panels in being sampled randomly rather than selected on the basis of recognized expertise. Given the large number of panels, this strategy was the only practical approach and reflects current practice in England where the National Institute of Clinical Excellence seeks participation of practicing health professionals. Our participation rates were low but as expected, based on advice from the Medical Research Council GP Research Framework. Participating GPs were typical of GPs in England with respect to age and sex distribution, but may differ from nonparticipants in other ways just as people who participate in guideline development groups may differ from those who are unwilling to take part.

Our findings are likely to be applicable to other conditions where the pathogenesis is unclear and psychosocial factors may contribute. However, they may not be generalizable to conditions with fewer psychosocial determinants and a clearer pathogenesis. Finally, we considered eighteen interaction tests at the $p < .05$ for inclusion in each model, without making any adjustment for multiple testing, so some caution is needed in interpreting these findings.

Implications of Results

In terms of the individual factors considered, this study supports the contention that participants' specialty can be an important influence on their judgments about appropriate treatment. The differences we found between men and women and by age, although statistically significant, are unlikely to have material effects other than at the margin.

CONTACT INFORMATION

James Carpenter, DPhil (james.carpenter@lshtm.ac.uk), Senior Lecturer, Department of Epidemiology and Population Health, **Andrew Hutchings**, MSc (andrew.hutchings@lshtm.ac.uk), Lecturer, Department of Public Health and Policy, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK

Rosalind Raine, PhD (r.raine@ucl.ac.uk), Professor, Department of Epidemiology and Public Health, University College London, 1-19 Torrington Place, London WC1E 6BT, UK

Colin Sanderson, PhD (colin.sanderson@lshtm.ac.uk), Reader, Department of Public Health and Policy, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK

REFERENCES

1. Brook RH, Chassin MR, Fink A, et al. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care*. 1986;2:53-63.
2. Burgers JS, Grol R, Klazinga NS, Makela M, Zaat J, for the AGREE collaboration. Towards evidence-based clinical practice: An international survey of 18 clinical guideline programs. *Int J Qual Health Care*. 2003;15:31-45.
3. Chan BT, Austin PC. Patient, physician and community factors affecting referrals to specialists in Ontario, Canada: A population-based, multi-level modelling approach. *Med Care*. 2003;41:500-511.
4. Coulter I, Adams A, Shekelle P. Impact of varying panel membership on ratings of appropriateness in consensus panels: A comparison of a multi- and single disciplinary panel. *Health Serv Res*. 1995;30:577-591.

5. Goldstein H. *Multilevel statistical models*. London: Arnold; 1995.
6. Hutchings A, Raine R. A systematic review of factors affecting the judgments produced by formal consensus development methods in health care. *J Health Serv Res Policy.* 2006;11:172-179.
7. Hutchings A, Raine R, Sanderson C, Black N. An experimental study of determinants of the extent of disagreement within clinical guideline development groups. *Qual Saf Health Care.* 2005;14:240-245.
8. Hutchings A, Raine R, Sanderson C, Black N. A comparison of formal consensus methods used for developing clinical guidelines. *J Health Serv Res Policy.* 2006;11:218-224.
9. Kahan JP, Bernstein SJ, Leape LL, et al. Measuring the necessity of medical procedures. *Med Care*. 1994;32:357-365.
10. Kahan JP, Park RE, Leape LL, et al. Variations by specialty in physician ratings of the appropriateness and necessity of indications for procedures. *Med Care.* 1996;34:512-523.
11. McKinlay JB, Burns RB, Durante R, et al. Patient, physician and presentational influences on clinical decision making for breast cancer: Results from a factorial experiment. *J Eval Clin Pract*. 1997;3:23-57.
12. McKinlay JB, Lin T, Freund K, Moskowitz M. The unexpected influence of physician attributes on clinical decisions: Results of an experiment. *J Health Soc Behav.* 2002;43:92-106.
13. MLwiN [computer program]. Version 2.0. London: Institute of Education; 2000.
14. Murphy MK, Black NA, Lamping DL, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assess*. 1998;2:1-88.
15. Park RE, Fink A, Brook RH, et al. Physician ratings of appropriate indications for six medical and surgical procedures. *Am J Public Health*. 1986;76:766-772.
16. Raine R, Haines A, Sensky T, et al. Systematic review of mental health interventions for patients with common somatic symptoms: Can research evidence from secondary care be extrapolated to primary care? *BMJ*. 2002;325:1082-1085.
17. Raine R, Sanderson C, Hutchings A, et al. An experimental study of determinants of group judgments in clinical guideline development. *Lancet*. 2004;364:429-437.
18. Richardson FM. Peer review of medical care. *Med Care.* 1972;10:29-39.
19. Scott EA, Black N. Appropriateness of cholecystectomy in the United Kingdom: A consensus panel approach. *Gut*. 1991;32:1066-1070.
20. Taffé P, Burnand B, Wietlisbach V, Vader JP. Influence of clinical and economical factors on the expert rating of appropriateness of preoperative use of recombinant erythropoietin in elective orthopedic surgery patients. *Med Decis Making*. 2004;24:122-130.