Published in final edited form as:
Psychosomatics. 2011 ; 52(4): 346–353. doi:10.1016/j.psym.2011.01.012.

Variability of Judgments of Capacity: Experience of Capacity Evaluators in a Study of Research Consent Capacity

Scott Y. H. Kim[1], **Paul S. Appelbaum**[2], **H. Myra Kim**[3], **Ian F. Wall**[4], **James A. Bourgeois**[5], **Bernard Frankel**[6], **Kevin C. Hails**[7], **James R. Rundell**[8], **Kathleen M. Seibel**[9], and **Jason H. Karlawish**[10]

[1]Center for Bioethics and Social Sciences in Medicine and Department of Psychiatry, University of Michigan, Ann Arbor, MI

[2]Division of Law, Ethics, and Psychiatry, Department of Psychiatry, Columbia University and New York State Psychiatric Institute, New York, NY

[3]Center for Statistical Consultation and Research, University of Michigan, Ann Arbor, MI

[4]Center for Behavioral and Decision Sciences in Medicine, University of Michigan, Ann Arbor, MI

[5]Department of Psychiatry & Behavioural Neurosciences, Michael G. DeGroote School of Medicine, McMaster University and St. Joseph's Healthcare, Hamilton, Ontario Canada

[6]Department of Psychiatry, University of California, San Diego, CA

[7]Department of Psychiatry, Albert Einstein Medical Center, Philadelphia, PA

[8]Department of Psychiatry, Mayo Clinic, Rochester, MN

[9]Department of Psychiatric Medicine, Brody School of Medicine at East Carolina University, Greenville, NC

[10]Departments of Medicine and Medical Ethics, Division of Geriatrics, Alzheimer's Disease Center, Center for Bioethics, and the Leonard Davis Institute for Health Economics, University of Pennsylvania, Philadelphia, PA

Abstract

BACKGROUND—Assessment of decision-making capacity is a common and important function of psychiatric consultants. However, the sources of variability in evaluators' judgments have not been well characterized.

OBJECTIVE—To examine the degree and potential sources of variability in the categorical capacity judgments of experienced psychiatrists.

METHOD—The setting was a study comparing the decision-making capacities of 188 persons with Alzheimer's disease to appoint a research proxy and to consent to two hypothetical randomized controlled trials for dementia (a new drug RCT and a neurosurgical RCT). We

Corresponding author: Scott Kim, MD, PhD, 300 North Ingalls St, 7C27, Ann Arbor, MI 48109, scottkim@med.umich.edu, 734-936-5222, 734-936-8944 (fax).

Disclosures: Authors report no applicable financial conflicts of interest.

compared 5 experienced consultation psychiatrists' capacity judgments for 555 videotaped capacity interviews. Both quantitative and qualitative data were used.

RESULTS—Pairwise kappa statistics ranged from slight agreement (0.17) to substantial agreement (0.64) with group kappa statistics ranging from fair to moderate agreement (0.40 to 0.45) for the psychiatrists' judgments regarding the three capacities. The sources of variability included varying "strictness" among judges, moderate test-retest reliability within judges, the relative novelty of assessing decision-making capacity for research participation decisions, as well as the limitations of the methods used to obtain capacity judgments in the study.

DISCUSSION—There is considerable variability in capacity judgments of experienced consultation psychiatrists regarding the capacities to appoint a research proxy and to consent to research. The potential sources of variability identified in this study may provide starting points for more effective training in capacity assessment.

The study of decision-making capacity has grown steadily over the past three decades.1 Most of this research has focused on the nature and degree of decisional impairment associated with various clinical states, including psychiatric 2–5, neurologic 6–8, and general medical9–11 conditions. Although this research has produced valuable information on the ranges of impairment of patients with various conditions, considerably less attention has been paid to how evaluators use this information to arrive at a categorical capacity judgment, i.e., about whether a person does or does not have capacity.12

Among the few studies that report on the categorical judgments of evaluators, the results have been mixed, with some studies reporting high rates of disagreement in capacity judgments13–15, and others showing somewhat more agreement.4, 6 Since significant consequences follow from a judgment of incapacity, variability in judgments that depend on the evaluator is not ideal. There is a need to better understand the degree and sources of disagreement in capacity determinations so that remedial approaches might be identified.

We report here the experiences of 5 consultation psychiatrists asked to provide categorical capacity judgments on 555 videotaped semi-structured interviews designed to assess three different types of decision-making capacity. We examine the reliability of categorical capacity judgments among the five judges, explore potential explanations for the variability using both quantitative and qualitative data, and consider options for increasing the reliability of capacity determinations.

METHOD

Procedures

This study was part of an NIH-sponsored project comparing different types of decision-making capacities of 188 persons with Alzheimer's disease. The study examined the capacity to appoint a proxy decision-maker for research consent and the capacity to give informed consent for two hypothetical research studies of varying risks and benefits. The main results of that study are published elsewhere.16

We adapted the MacArthur Assessment Tool-Clinical Research (MacCAT-CR) for two research scenarios used in previous studies.15 One scenario consisted of a randomized clinical trial of a medication for Alzheimer's disease ("drug RCT") and the other consisted of a randomized placebo-controlled (sham surgery) neurosurgical trial of cell transplant intervention ("neurosurgical RCT"). The MacCAT-CR is structured according to the four-abilities model of decision-making capacity.17 These include "*understanding* [emphasis added] of disclosed information about the nature of the research project and its procedures;

appreciation of the effects of research participation (or failure to participate) on subjects' own situations; *reasoning* about participation; and ability to communicate a *choice*."18

To assess a patient's capacity to appoint a research proxy decision-maker (that is, capacity to appoint someone to make a decision in the subject's place regarding research participation), we developed the Capacity to Appoint a Proxy Assessment (CAPA). The CAPA follows a similar four abilities (understanding, appreciation, reasoning, and evidencing a choice) framework to MacCAT-CR, using similar scoring criteria. Details of the CAPA are presented elsewhere.16

Although the MacCAT-CR and the CAPA each provide a score for understanding, appreciation, reasoning, and choice, they do not yield a categorical capacity decision. That decision requires a clinician's judgment. In this study, the criterion standard for capacity status was the majority (or greater) agreement of 5 experienced consultation psychiatrists' capacity judgments. The psychiatrists were recruited from the membership of the Academy of Psychosomatic Medicine (APM), the primary professional society for consultation psychiatrists, a subgroup of psychiatrists who are most experienced in capacity determinations in the clinical setting. Our judges had on average 29.4 (SD 8.3) years of post-residency or fellowship clinical experience, represented different parts of the United States (West Coast, Southeast, Midwest, Northeast), had a range of current rate of performing capacity evaluations with an estimated average of 109.8 clinical capacity evaluations per year (SD 139.3; one judge was nearing retirement with current yearly rate of 9). The expert judges were trained using PowerPoint® presentations and 5 practice interviews, with two one-hour conference calls to explain their task, review their experience and answer questions regarding the practice interviews. They rendered their capacity judgments independently of one another, basing their judgments solely on each videotaped capacity interview.

Each judge rendered judgments for every interview, for a total of 555 judgments distributed over 36 months between the years 2006-2009. (Of the 188 AD subjects, 7 refused one of the MacCAT-CR interviews and 1 refused both MacCAT-CR interviews.) The monthly batches consisted of approximately 16 interviews each (5-6 of each interview type). The review schedule used a stratified random sample based on patient MMSE, selecting at random within three groups according to MMSE scores: 17 or below, 18-23, 24 or higher; time between reviewing interviews of the same patient averaged 8.2 months (SD 4.7). As a quality control measure, the judges participated in conference calls led by the first author at roughly 6-month intervals to discuss three interviews (one each of the three types of capacity interviews randomly chosen among those without unanimous agreement in order to generate discussion) from the monthly batch they had just rated.

The judges rendered their determinations by filling out a form that asked: "Based on the interview of the subject you just saw, do you believe that this person has sufficient capacity to give his/her own informed consent to the research study?" (or "to appoint a proxy for research decisions?" depending on the interview). The four potential responses (which we will refer to as "capacity scores") were 1=definitely has sufficient capacity, 2= probably has sufficient capacity, 3= probably does not have sufficient capacity, and 4= definitely does not have sufficient capacity. Dichotomous judgments (patient has capacity vs. patient does not have capacity) were created by collapsing probable and definite ratings.12 The criterion standard for the final categorical capacity status for each subject was based on the agreement of three or more judges on dichotomous judgments. Finally, for each video, the expert judges were asked to rate the following: "The videotaped interview gave me sufficient basis to make my decision in this case," measured on a 5-point scale (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree).

Human Research Subjects

The research protocol was reviewed and approved by the IRBs at the University of Michigan, Michigan State University, and the University of Pennsylvania. Given the low level of risk of this interview study, the AD subjects provided their own consent when determined to be capable by the interviewer; otherwise, a surrogate gave permission in addition to subject assent.

Analysis

Analyses were conducted using STATA 8.0. We first assessed the reliability of the five-expert panel as the criterion standard for final capacity status by calculating the Cronbach's alpha, treating the five-expert panel's dichotomized judgments as five-items giving a summary capacity determination. This statistic reflects the reliability of the 5 experts when their assessments are summarized *as a group* to arrive at the capacity status of the AD subjects.

To examine the reliability of *individual* judgments, we first examined pair-wise comparisons of expert judges' dichotomized judgments. Agreement among all ten possible pairs of the five individual expert judges was assessed by calculating percent agreement and, to account for chance agreement, the kappa statistic.19 We also calculated Fleiss's group kappa statistic which gives a summary measure of the reliability of individual judgments for the five judges.20 Both pairwise and group kappa statistics reflect the reliability of individual expert judgments.

To assess potential explanations for variability in expert judgments, we examined the relative "strictness" of the judges in three steps. We first compared the means of the capacity scores (capacity judgments given on the 1-4 scale) across the 5 experts using analysis of variance. We then examined the correlation (Pearson correlation and R^2) between the pairwise differences in mean capacity scores and the pairwise kappa statistic, for all ten combinations of expert pairs to assess if greater threshold differences are associated with less agreement in judgments between expert pairs. Finally, we examined whether controlling for the variability in thresholds (as reflected in the mean capacity scores) would result in increased reliability of individual judgments by calculating the intra-class correlation coefficient (ICC) after accounting for expert effect. This was done using a linear mixed effects model where experts were treated as fixed effects and subjects as random effects. The resulting ICC without the expert effect corresponds to the group kappa statistic, and the ICC adjusting for expert effects corresponds to the expected group kappa adjusted for threshold differences. Adjusting for the expert effect essentially removes the extent of variability due to differences in experts' thresholds.

Lastly, we examined within expert "test-retest" reliability for 23 capacity judgments (8 CAPA, 7 drug RCT, and 8 neurosurgery RCT) for which the experts provided two separate judgments, by evaluating the same videotapes an average of 19.5 months apart (SD 11.0).

We analyzed two sources of qualitative data to explore sources of judgment variability. First, at the end of the project, the judges were asked to provide a narrative reflection on their experience by responding to the following questions: "What principles or practices have you found yourself following? Developing? Adapting or changing? What insights have you gained that could only have come from doing a lot of [capacity determinations]—i.e., what has experience taught you that goes beyond what one could anticipate? What makes the task challenging, and how do you deal with those points of challenge?"

Second, the research team and the 5 expert judges held a one-day retreat in 2010 during which the judges' experiences were discussed in detail. Using both notes and an audio

recording of the retreat, the PI (SK) and an assistant (IFW) prepared a summary of the main themes, which were then reviewed by the 5 judges and the other research team members (PSA, JHK).

RESULTS

Table 1 shows the capacity status of the subjects for each of the three capacities, along with the level of agreement in the 5 experts' judgments. The Cronbach's alpha was 0.80 for the determination of capacity to appoint a research proxy, 0.85 for determination of the capacity to consent to the drug RCT, and 0.81 for determination of the capacity to consent to the neurosurgical RCT. These alpha values can be interpreted as the expected correlation between capacity judgments based on two different random samples of five-expert panels and thus indicate that the reliability of using a five-expert panel as the criterion standard for determining capacity status is quite high.[21, 22]

Table 2 compares the mean capacity scores given by the 5 judges on the 1-4 scale (1=definitely has capacity, 2=probably has capacity, 3=probably does not have capacity, and 4=definitely does not have capacity). The mean capacity scores of the experts are significantly different for each of the three capacities. Further, across the three capacities, the relative "strictness" of the judges appears to be preserved.

Table 3 presents the agreement of experts' capacity judgments for each of the 10 possible pairs of judges, given as both percent agreement and as kappa statistics. One often-cited guideline for interpretation of the kappa statistic is from Landis and Koch[23], in which kappa<0.00 indicates poor agreement, 0.00-0.20 indicates slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 almost perfect agreement. Using this guideline, for judgments of capacity to appoint a proxy, 5 of the 10 expert judge pairs had "fair" agreement and the other 5 pairs had "moderate" agreement. For judgments on the capacity to consent to drug RCT, 3 of 10 pairs had "fair" agreement, 4 pairs had "moderate" agreement, and 3 pairs had "substantial" agreement. For the capacity to consent to the neurosurgical RCT, 1 of 10 pairs had only "slight" agreement, 3 pairs "fair" agreement, 6 pairs "moderate," and 1 pair "substantial" agreement.

Table 3 also shows the pair-wise mean differences between experts in the capacity scale scores. The correlation coefficient of the relationship between these mean differences with the corresponding pairwise kappa statistics was −0.77 (R^2 = 0.59; p=.009) for the determinations of capacity to appoint a proxy, −0.73 (R^2 = 0.53; p=.02) for the capacity to consent to drug RCT, and −0.87 (R^2 = 0.76; p=.001) for the capacity to consent to neurosurgical RCT, suggesting that differences in capacity thresholds (i.e., how 'strict' or 'lenient' the judges are) between experts (shown by the mean difference in capacity scores) account for a significant amount of variability in capacity judgments among the judges. The expected group kappa statistics, when adjusted for expert capacity scores, increased from 0.40 to 0.44 (p<0.001) for the determination of the capacity to appoint a research proxy, from 0.45 to 0.53 (p<0.001) for the determination of the capacity to consent to the drug RCT, and from 0.41 to 0.46 (p<0.001) for the determination of the capacity to consent to the neurosurgical RCT.

Qualitative Data

The results of the qualitative analysis of the experts' comments at the end of the project period and of the comments from the one-day retreat revealed several potential sources of variability as well as other points of interest. The main themes can be grouped under three categories.

First, there were issues specific to the methods used in our study. The judges felt constrained by the somewhat artificial conditions under which they rendered their categorical judgments since they did not conduct the interviews themselves and did not have other background information usually available in clinical settings (such as results of cognitive tests). The judges also noticed that some interviewers (i.e., research assistants who administered the capacity interviews) in the videotapes were better interviewers than others—for example, showing more empathy and patience and more appropriate probing to clarify ambiguous responses. Thus, some capacity judgments had to be based on less than optimal information and some degree of "filling in the gaps" or interpretation had to be supplied by the experts, introducing a potential source of variability.

Another methodological issue derived from the fact that in clinical practice if a patient refuses a proposed treatment, the capacity assessment is an assessment of whether the patient is *capable of refusing* a particular treatment, whereas for the purposes of this project, we asked our expert judges to render a judgment regarding whether the person on the videotape was *capable of consenting* to an RCT (or appointing a proxy), even if the subject said he or she would not want to be in the RCT (or did not want to appoint a proxy). Thus, some variability may have arisen due to this forced "as if" judgment that we required of the experts.

It is possible that our experts handled these demands for interpretation in varying ways, contributing to the inconsistency in their judgments. Supporting this possibility is that there was a range in the experts' perceptions regarding the sufficiency of the videotaped interviews as the basis for their judgments. In rating the statement, "The videotaped interview gave me sufficient basis to make my decision in this case," (using a 5-point scale: 1=strongly agree, 2=agree, 3=neither agree nor disagree, 4=disagree, 5=strongly disagree), mean scores among the judges had a range of 1.7-2.9 for the capacity to appoint a proxy interviews, 1.6-2.7 for the drug RCT interviews, and 1.2-2.6 for the neurosurgical RCT interviews.

The second category of themes concerned the relative novelty of the capacities that the experts were asked to evaluate. The capacity to appoint a research proxy, for instance, has never been systematically studied. At least one judge commented that his categorical judgments regarding this capacity evolved over the course of the 36 months of the study. This judge began to realize over time that many subjects, despite their obvious cognitive impairment, genuinely expressed their desire to help others and to contribute to research; he came to feel that applying too high a threshold for determining the capacity to appoint a proxy would make it difficult to honor such desires. This judge's mean capacity score (for the capacity to appoint a research proxy) for the first 20% of cases was 3.1, but the final 20% of cases had a mean score of 1.7, indicating that he did indeed make a change in his judgments over time. In fact, there was other evidence of some degree of variability within experts over time. The test-retest analysis of 23 capacity judgments (8 capacity to appoint a proxy, 7 capacity to consent to drug RCT, and 8 capacity to consent to neurosurgical RCT) revealed within-judge kappa statistics ranging from 0.23 ("fair" agreement) to .71 ("substantial" agreement).

Third, there were conceptual issues for which we do not yet have widely accepted guidance that may have increased variability. One recurring question among the expert judges was: How long must the person retain the information to be deemed to have intact capacity? This was a particularly relevant question for our study given that short term memory loss is one of the earliest signs of Alzheimer's disease. The lack of a widely accepted answer to this question may have contributed to variability in judgment. Another recurring theme among the judges was: Do the subjects show that they grasp "the big picture?" As one judge put it,

"if you don't understand the concept, the details won't save you." Although all of the judges felt that this was an important criterion, what constitutes "grasping the big picture" may have varied among the judges since there are currently no uniform guidelines on how to use such a criterion and since the semi-structured capacity interviews used in this study were not specifically designed to highlight this concept.

DISCUSSION

Capacity determinations rely heavily on individual clinician judgments guided by fairly broad standards, such as understanding, appreciation, reasoning, and choice.1 Yet how clinicians arrive at their categorical judgments of capacity remains poorly understood. Previous research by Marson and colleagues14 showed that five experienced clinicians (from geriatric psychiatry, neurology, and geriatric medicine) evaluating the capacity of persons with mild AD and normal elderly achieved only 56% agreement. In an experimental video survey of 99 consultation psychiatrists using the same two RCT scenarios used for the present study, we found that the subject portrayed in the drug RCT video for that experiment was deemed competent by 40% of the psychiatrists, whereas the remainder believed that he was incompetent to provide informed consent.15 Multivariate models did not identify any characteristics of the psychiatrists that predicted their judgments.

In the present study, we explored the degree and potential sources of variability among 5 consultation psychiatrists serving as expert judges providing categorical capacity judgments. Although we found that a 5-judge *panel* did have considerable reliability, it is not realistic to use a 5-person panel routinely. The examination of reliability of the experts' judgments as individuals revealed widely ranging levels of agreement, with an average non-chance agreement in the "moderate" range. A significant source of variability appeared to be due to the judges employing different thresholds, with some judges being "stricter" than others in their judgments, as evidenced by significant differences in their mean capacity scores. Further, similarity in capacity thresholds was highly correlated with the level of agreement among pairs of judges. When the group kappa statistic was adjusted for the mean capacity score differences in a linear mixed effects model, the agreement levels increased signficantly. Although this increase in kappa was only modest, it is notable that all three ways of looking at this issue yielded consistent results. Another significant source of variability was within-expert reliability, as our test-retest reliablity check resulted in kappa statistics of 0.23 to 0.71, indicating that the variation within judges was comparable to the variability between judges. This may explain the estimated modest increase in group kappa statistics even after removing variability due to expert-specific thresholds.

What are the implications? First, at present, the determination of capacity to provide consent to research, although increasingly studied, is still a relatively new area of clinical practice. Thus, it should not be surprising that there is considerable variability among evaluators. Until the practice becomes more widespread with shared guidelines and experiences, policies regulating research consent should be sensitive to this potential for variability among individual evaluators. This may mean that for high stakes situations—such as research protocols involving significant risk to subjects—consideration should be given to providing sufficient resources (e.g., multi-member panels of experienced evaluators using validated interviews) to ensure reliability and validity of capacity judgments.

Second, there are implications for training of capacity evaluators. One previous survey showed that members of the Academy of Psychosomatic Medicine received relatively little training in capacity evaluation, with only 2.6 lectures and 3 supervised cases during their training.15 Moreover, our experience of clinical practice indicates that a capacity evaluator does not usually compare his or her judgments of capacity with other evaluators. Thus, it is

possible that the high variability in capacity judgments may be due to lack of standardized training. However, as the need for capacity evaluations in the general hospital grows with an aging population having a high prevalence of decisional impairment, as well as with an increasing number of research studies involving the decisionally impaired, a more systematic and transparent method of training may be needed. There is evidence that specific guidance given to evaluators can increase reliability.24 Further, given the data that capacity evaluators may use differing thresholds, a focus on "calibrating" these thresholds across capacity evaluators should be a key consideration.

There are, however, limitations to our study and the results should be interpreted with caution. First, all of the capacity judgments studied were for the research consent (or research proxy appointment) context, which is still a relatively new area of practice. Thus, generalizing to the treatment context should be done cautiously. However, insofar as the issues raised in our report are relevant to the clinical setting—e.g., the relative "strictness" of evaluators—our findings may shed light even on evaluations of capacities outside the research consent context. Second, as revealed in our qualitative analyses, some of the variability may have been due to artifacts of the design of our study that may not be present in the ordinary practice context. Specifically, our judges did not conduct their own interviews which would have allowed them to conduct individualized probing of unclear areas (rather than just viewing a videotaped interview conducted by research staff) and they did not have the kind of background clinical information they otherwise might have had. However, it is also possible that more individualized styles of interviewing with less standardization could have led to even more variability in judgment. Third, although we attempted to recruit our experts from a variety of geographical regions, they cannot be seen as representative of all psychiatrists. They were, however, highly experienced as capacity evaluators.

We close with two specific issues raised by our expert judges in the course of their review of over 500 interviews, regarding how to apply the standards for assessing capacity. First, in the course of applying the understanding standard, our expert judges raised an issue that is not explicitly measured in the MacCAT-CR instrument, namely, how long a patient must retain information to be deemed to meet the understanding standard. The recent Mental Capacity Act 2005 (England and Wales)25 explicitly recognizes that indefinite retention is not necessary. But how should a clinician apply this concept in practice? The most obvious interpretation is that the subject must retain key information long enough to make the decision at hand. Whether this is sufficient may require further discussion in the field and clarification by the courts. Second, we found the concept of "getting the big picture" intriguing and important. Currently, this is not a concept that is widely taught in the evaluation of capacity, although it may reflect how many evaluators reach their judgments. Where this approach fits into the now-familiar four abilities model of capacity26 and how to operationalize it remains unclear. However, given the importance that our expert judges placed on the "big picture," clarifying how to apply the concept may be an important way of increasing the reliability and validity of capacity determinations in general.

Acknowledgments

Supported by NIH R01-MH075023.

References

1. Kim, SYH. Evaluation of Capacity to Consent to Treatment and Research. Oxford University Press; New York: 2010.

2. Grisso T, Appelbaum PS. The MacArthur Treatment Competence Study. III: Abilities of Patients to Consent to Psychiatric and Medical Treatments. Law & Human Behavior. 1995; 19:149–74. [PubMed: 11660292]
3. Carpenter WT Jr. Gold JM, Lahti AC, et al. Decisional capacity for informed consent in schizophrenia research. Archives of General Psychiatry. 2000; 57:533–8. [PubMed: 10839330]
4. Kim SYH, Appelbaum PS, Swan J, et al. Determining when impairment constitutes incapacity for informed consent in schizophrenia research. The British Journal of Psychiatry. 2007; 191:38–43. [PubMed: 17602123]
5. Lapid MK, Rummans TA, Poole KL, et al. Decisional capacity of severely depressed patients requiring electroconvulsive therapy. Journal of ECT. 2003; 19:67–72. [PubMed: 12792453]
6. Kim SYH, Caine ED, Currier GW, Leibovici A, Ryan JM. Assessing the competence of persons with Alzheimer's disease in providing informed consent for participation in research. American Journal of Psychiatry. 2001; 158:712–7. [PubMed: 11329391]
7. Marson DC, Cody HA, Ingram KK, Harrell LE. Neuropsychological predictors of competency in Alzheimer's disease using a rational reasons legal standard [comment]. Archives of Neurology. 1995; 52:955–9. [PubMed: 7575222]
8. Dymek MP, Atchison P, Harrell L, Marson DC. Competency to consent to medical treatment in cognitively impaired patients with Parkinson's disease. Neurology. 2001; 56:17–24. [PubMed: 11148230]
9. Moser DJ, Schultz SK, Arndt S, et al. Capacity to Provide Informed Consent for Participation in Schizophrenia and HIV Research. American Journal of Psychiatry. 2002; 159:1201–7. [PubMed: 12091200]
10. Grisso T, Appelbaum PS, Hill-Fotouhi C. The MacCAT-T: a clinical tool to assess patients' capacities to make treatment decisions. Psychiatric Services. 1997; 48:1415–9. [PubMed: 9355168]
11. Palmer BW, Dunn LB, Appelbaum PS, et al. Assessment of Capacity to Consent to Research Among Older Persons With Schizophrenia, Alzheimer Disease, or Diabetes Mellitus: Comparison of a 3-Item Questionnaire With a Comprehensive Standardized Capacity Instrument. Archives of General Psychiatry. 2005; 62:726–33. [PubMed: 15997013]
12. Kim SYH. When Does Decisional Impairment Become Decisional Incompetence? Ethical and Methodological Issues in Capacity Research in Schizophrenia. Schizophrenia Bulletin. 2006; 32:92–7. [PubMed: 16177276]
13. Marson DC, Hawkins L, McInturff B, Harrell LE. Cognitive models that predict physician judgments of capacity to consent in mild Alzheimer's disease. Journal of the American Geriatrics Society. 1997; 45:458–64. [PubMed: 9100715]
14. Marson DC, McInturff B, Hawkins L, Bartolucci A, Harrell LE. Consistency of physician judgments of capacity to consent in mild Alzheimer's disease. Journal of the American Geriatrics Society. 1997; 45:453–7. [PubMed: 9100714]
15. Kim SYH, Caine ED, Swan JG, Appelbaum PS. Do clinicians follow a risk-sensitive model of capacity determination? An experimental video survey. Psychosomatics. 2006; 47:325–9. [PubMed: 16844891]
16. Kim S, Karlawish J, Kim HM, Wall IF, Bozoki A, Appelbaum PS. Preservation of the capacity to appoint a proxy decision maker: implications for dementia research. Arch Gen Psychiatry. Accepted for publication.
17. Grisso, T.; Appelbaum, PS. Assessing Competence to Consent to Treatment: A guide for physicians and other health professionals. Oxford University Press; New York: 1998.
18. Appelbaum PS, Grisso T, Frank E, O'Donnell S, Kupfer D. Competence of depressed patients for consent to research. American Journal of Psychiatry. 1999; 156:1380–4. [PubMed: 10484948]
19. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and psychological measurement. 1960; 20:37–46.
20. Fleiss, JL.; Levin, BA.; Paik, MC. Statistical methods for rates and proportions. 3rd ed.. J. Wiley; Hoboken, N.J.: 2003.
21. Streiner, DL.; Norman, GR. Health measurement scales. Oxford University Press; New York: 2003.

22. Thorndike, R. Applied psychometrics. Houghton Mifflin; Boston: 1982.

23. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977; 33:159–74. [PubMed: 843571]

24. Marson DC, Earnst KS, Jamil F, Bartolucci A, Harrell L. Consistency of physicians' legal standard and personal judgments of competency in patients with Alzheimer's disease. Journal of the American Geriatrics Society. 2000; 48:911–8. [PubMed: 10968294]

25. Mental Capacity Act. [Accessed September 28, 2010]. 2005 at http://www.opsi.gov.uk/acts/acts2005/50009--b.htm#30.

26. Appelbaum PS. Assessment of Patients' Competence to Consent to Treatment. New England Journal of Medicine. 2007; 357:1834–40. [PubMed: 17978292]

Table 1

Capacity status of persons with Alzheimer's disease, for three different decision-making tasks, as determined by a 5-expert panel

	Capacity to Appoint Proxy (n=188)*	Capacity to Consent to Drug RCT (n=181)*	Capacity to Consent to Neurosurgical RCT (n=186)*
	N (%)	N (%)	N (%)
Capacity	116 (61.7)	75 (41.4)	29 (15.6)
3 judges agree	32 (17.0)	20 (11.0)	8 (4.3)
4 judges agree	27 (14.4)	30 (16.6)	14 (7.5)
5 judges agree	57 (30.3)	25 (13.8)	7 (3.8)
No capacity	72 (38.3)	106 (58.6)	157 (84.4)
3 judges agree	23 (12.2)	21 (11.6)	19 (10.2)
4 judges agree	26 (13.8)	29 (16.0)	34 (18.3)
5 judges agree	23 (12.2)	56 (30.9)	104 (55.9)

* The numbers are different because 188 completed the first interview which included the CAPA and either the drug RCT or the neurosurgical RCT MacCAT-CR (randomly chosen), 7 declined the second MacCAT-CR interview, and one subject finished neither of the MacCAT-CR interviews.

Table 2

Expert judges' mean capacity scores* for the three capacities

	Expert Judge					ANOVA (p-value, F)
	1	2	3	4	5	
Capacity to Appoint a Proxy(n=188)	2.45 (1.11)	1.73 (0.91)	2.49 (1.00)	2.25 (0.97)	2.63 (1.17)	<0.001, 21.6
Capacity to Consent to Drug RCT, (n= 181)	3.29 (0.94)	2.15 (1.17)	2.69 (1.04)	2.71 (1.03)	3.10 (1.09)	<0.001, 31.3
Capacity to Consent to Neurosurgery RCT, (n=186)	3.78 (0.62)	2.76 (1.18)	3.33 (0.86)	3.09 (0.93)	3.52 (0.80)	<0.001, 35.6

*Scores: 1=Definitely has capacity, 2=Probably has capacity, 3=Probably does not have capacity, 4=Definitely does not have capacity.

Table 3

Pair-wise comparisons of expert judgments for three types of capacity

	Capacity to Appoint a Proxy (n=188)			Capacity to Consent to Drug RCT (n=181)			Capacity to Consent to Neurosurgery RCT (n=186)		
Expert Judge Pair	**Kappa**	**% agreement**	**Capacity score mean difference[1]**	**Kappa**	**% agreement**	**Capacity score mean difference[1]**	**Kappa**	**% agreement**	**Capacity score mean difference[1]**
1 – 2	0.40	72	0.72**	0.22	54	1.14**	0.17	67	1.02**
1 – 3	0.58	79	−0.04	0.39	70	0.60**	0.36	86	0.45**
1 – 4	0.51	76	0.20*	0.44	75	0.58**	0.29	81	0.70**
1 – 5	0.36	68	−0.18	0.42	77	0.18*	0.56	92	0.26**
2 – 3	0.33	67	−0.76**	0.61	80	−0.54**	0.49	78	−0.57**
2 – 4	0.39	73	−0.52**	0.45	71	−0.56**	0.43	75	−0.32**
2 – 5	0.23	62	−0.89**	0.38	66	−0.96**	0.37	74	−0.76**
3 – 4	0.47	73	0.25**	0.61	81	−0.02	0.48	83	0.25**
3 – 5	0.50	75	−0.13	0.62	81	−0.41**	0.64	90	−0.19**
4 – 5	0.31	65	−0.38**	0.56	80	−0.39**	0.50	85	−0.44**
Group	0.40	na	na	0.45	na	na	0.41	na	na

* paired t-test significant at .01 level

** paired t-test significant at .001 level.

[1] Score range: 1= Definitely has capacity, 2=Probably has capacity, 3=Probably does not have capacity, 4=Definitely does not have capacity