Face Validity and Decision Aid Neglect

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

James E. Kajdasz, M.A.

Graduate Program in Psychology

The Ohio State University

2010

Dissertation Committee:

Dr. Hal R. Arkes, Advisor

Dr. Thomas E. Nygren

Dr. Michael L. DeKay
The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.
Abstract

Test makers are generally not concerned with how suitable a test item is perceived to be by the testee (face validity). Unlike test takers however, decision makers (DMs) may choose to disregard a decision aid if they believe it is unsuitable for its stated purpose. I present a scale adapted from Nevo (1985) to measure decision aid face validity (DAFV).

In three experiments, measures of DAFV obtained from one group of participants are shown to correlate with rates of decision aid reliance in another group of participants. Participants were asked to estimate the rent of an apartment. After making an initial estimate for the apartment, the participant was then presented with one of several possible DAs designed to estimate rent. The methodology of the aid was described along with the DA estimated rent. Participants were given the option to revise their initial rent estimate based on the new information provided by the DA if they desired. The FV of the aid (as measured during a previous pilot study) was correlated with the measure of DA reliance.

I also present research that integrates DAFV with decision maker confidence. The previous methodology is repeated except the apartments are located in four cities. Participants have high confidence in their ability to estimate the rent in the local city, but less confidence in their ability to estimate rent in far away cities. As shown in previous research, a negative correlation was observed between confidence and decision aid
reliance. In addition, a hypothesized interaction was observed between confidence and
DAFV. When participant confidence was high (i.e., local city), DA reliance tended to be
low. When confidence was low (i.e. foreign city), DA reliance was high if DAFV was
high. When DAFV was low, DA reliance tended to be low even when confidence was
low.

Finally, how the performance of a DA is communicated was examined. The
performance of a hypothetical decision aid was communicated in several different
formats: 1) participants saw the mean accuracy of the DA, and the mean accuracy of the
DM: “The decision aid was, on average, 70% accurate compared to 60% accuracy of
unaided decision makers.” 2) participants were given mean DA accuracy only “The
decision aid was, on average, 70% accurate.” 3) Participants were given the proportion of
DMs beaten by the DA: “This decision aid performed better than 76% of human resource
experts who did not use the aid.” Face validity was assessed for the aid using the various
formats. Two experiments replicate the finding that “DA only” had consistently higher
FV than “DA & DM” performance. This occurred even though the DA always
performed better than the DMs. It seems participants were unimpressed by an aid that
only outscored its expert decision maker by 10% or less, and punished the DA when this
difference was made salient.
For my son.

I hope it might be significant in some way, but this dissertation will never be as important as your crayon drawings of monster trucks.
Acknowledgments

We all play a balancing act in our lives. Demands placed on us typically exceed the time we have to meet them. The more selfless we are, the more demands we seem to acquire. Academic advisors are no different. As a graduate student, we can sometimes only grab the occasional attention of our mentors. This has not been the case with my advisor, Dr. Hal Arkes. I honestly don’t know how he does it. He appears to be constantly engaged in a ubiquitous cacophony of authoring book chapters/authoring articles/research meetings/student meetings… Through all of it, I always felt like I remained a priority and never wanted for guidance or attention. Most of us strike a balance but a few of us appear to have figured the code to be both selfless and still meet and exceed obligations. What a great pleasure and personal example it has been to work with Hal. Thank you, sir.
Vita

1996.................................................................B.S. Behavioral Science, United States Air Force Academy

2003.................................................................M.A. Psychology, George Mason University

1996 to Present..................................................Intelligence Officer, United States Air Force

Fields of Study

Major Field: Psychology
## Table of Contents

Abstract ............................................................................................................................... ii  

List of Tables ................................................................................................................... viii  

List of Figures .................................................................................................................... ix  

Chapter 1: Ignoring Good Advice: The Tragedy of Decision Aid Neglect ................. 1  

Chapter 2: Measuring Decision Aid Face Validity ......................................................... 10  

   Study 1: Measuring Face Validity of Decision Aids ..................................................... 10  

Chapter 3: Using Face Validity to Predict Decision Aid Neglect ............................... 17  

   Study 2: Face Validity and Decision Aid Neglect ....................................................... 17  

   Study 3: DA Procedure as a Determinant of Face Validity and DA Neglect .......... 28  

Chapter 4: Additional Variables: Confidence, Face Validity and DA Neglect .......... 38  

   Study 4: Obtaining Measures of Confidence and Decision Aid Face Validity ...... 39  

   Study 5: Confidence, Face Validity, and Decision Aid Neglect ............................ 44  

Chapter 5: Influencing Decision Aid Face Validity ...................................................... 66  

   Study 6: Communicating DA Performance and its Effect on DA Face Validity .... 69  

   Study 7: Communicating DA Performance and Face Validity (2nd Experiment) .... 83  

Chapter 6: General Discussion ....................................................................................... 93  

References ....................................................................................................................... 101  

vii
List of Tables

Table 1. Test takers versus decision makers................................................................. 9
Table 2. Study 2. Predictors of participant final rent estimate........................................ 24
Table 3. Study 3. Predictors of participant final rent estimate........................................ 34
Table 4. Study 5. DAFV as a predictor of participant final rent estimate ....................... 50
Table 5. Study 5. Confidence as a predictor of final rent estimate................................. 54
Table 6. Study 5. Predictors of weight-of-advice ........................................................... 61
Table 7. Study 5. DAFV & confidence as predictors of final rent estimate ................. 63
Table 8. Study 6. Predictors of decision aid face validity ............................................. 81
Table 9. Study 7. Mean face validity of decision aid..................................................... 89
Table 10. Study 7. Predictors of decision aid face validity........................................... 90
List of Figures

Figure 1. Study 1. FV of various methods to estimate rent (relative method).............. 14
Figure 2. Study 1. FV of various methods to estimate rent (absolute rating method).... 16
Figure 3. Study 2. Example apartment data viewed by participants.............................. 19
Figure 4. Study 2. Histogram of participant correlations (DAFV vs. WOA)............... 23
Figure 5. Study 2. Predicted final rent: Interaction plot (DA Advice * DAFV)......... 27
Figure 6. Study 3. Histogram of participant correlations (DAFV vs. WOA)............. 33
Figure 7. Study 3. Predicted final rent: Interaction plot (DA Advice * DAFV)......... 36
Figure 8. Hypothesized relationship between WOA, Confidence, and DAFV .......... 39
Figure 9. Study 4. Participants’ confidence in estimating rent in various cities......... 41
Figure 10. Study 4. FV of methods to estimate rent (absolute rating method)......... 42
Figure 11. Study 4. Rent estimation methods whose FV does not vary by city......... 44
Figure 12. Study 5. Histogram of participant correlations (DAFV vs. WOA).......... 49
Figure 13. Study 5. Predicted final rent: Interaction plot (DA Advice * DAFV)...... 51
Figure 14. Study 5. Mean WOA for various confidence levels................................. 52
Figure 15. Study 5. Histogram of participant correlations (Confidence vs. WOA) .... 53
Figure 16. Study 5. Predicted final rent: Interaction plot (DA Advice * Conf) ......... 55
Figure 17. Study 5. Mean WOA (all values used) for varying Conf. and DAFV ...... 59
Figure 18. Study 5. Mean WOA (proper values only) for varying Conf. & DAFV .... 59
Figure 19. Study 5. Mean WOA (improper values converted) by Conf. & DAFV.............. 60
Figure 20. Study 5. Predicted WOA for varying confidence and DAFV ......................... 62
Figure 21. Study 5. Predicted final rent: Int. plot(DA Adv * DAFV * Conf).................. 64
Figure 22. Hypothesized FV when variance of DM performance changes.................... 68
Figure 23. Hypothesized relationship between DAFV and performance description. .... 70
Figure 24. Study 6. DAFV for ‘DA only’ and ‘DA&DM’ conditions ............................. 73
Figure 25. Study 6. DAFV for different performance conditions............................... 75
Figure 26. Study 6. DAFV for ‘DA only’ & ‘DA&DM’ conditions (by accuracy) ......... 78
Figure 27. Study 6. DAFV for ‘DA only’ & ‘DA&DM’ cond. (by DA-DM gap)............ 79
Figure 28. Study 6. Pred. DAFV int. plot (DA-DM gap vs. Accuracies presented). ...... 82
Figure 29. Study 7. Difference in FV between cond. (‘DA only’ – ‘DA & DM’). ......... 86
Figure 30. Study 7. DAFV for ‘DA only’ and ‘DA&DM’ conditions ............................. 87
Figure 31. Study 7. DAFV for ‘DA only’ & ‘DA&DM’ cond. (by DA-DM gap).......... 88
Figure 32. Study 7. Pred. DAFV int. plot (DA-DM gap vs. Accuracies presented) ..... 91
Chapter 1: Ignoring Good Advice: The Tragedy of Decision Aid Neglect

Decision aids have been developed in a variety of applications to increase the quality and accuracy of our judgment. Their potential to improve our decisions has been well established. A meta-analysis by Grove, Zald, Lebow, Snitz, and Nelson (2000) evaluated research in which expert decision makers were compared to statistical decision aids. The decision aids were applied in a wide variety of fields to include predictions of college academic performance, cerebral impairment, criminal recidivism, suicide attempts, job performance, success in military training, magazine advertising sales, coupon redemption, job turnover, and business failure. In 136 aids tested, the mechanical aids did at least as well as or outperformed clinical experts 128 times (94%). The mechanical aids were substantially better than experts about half the time. What wonderful potential exists if our society were to rely more on such aids, and less on subjective judgment (Baron, Bazerman, & Shonk, 2006)!

Despite such potential, decision makers are typically reluctant to rely on a decision aid over their own judgment or the subjective judgment of an expert. Examples of decision aid neglect are common even in the most critical settings, such as hospitals. Certain hospital patients are particularly susceptible to infection. These include patients in the pediatric intensive care unit (PICU), the emergency room (ER) and the bone marrow transplant unit (BMT). In an attempt to improve patient care, Rocha et al. (2001)
introduced a system designed to identify to caregivers those patients who were at highest risk, and who might require additional screening. The Computerized Pediatric Infection Surveillance System (COMPISS) provided an alert on the computer terminal next to the patient. Once the alert was acknowledged by the caregiver, information about the potential for infection, possible treatment, and notification procedures was provided. Clinical decision makers resisted being trained on the system, and in the end, COMPISS failed to modify caregiver behavior. The computer alerts frequently went unacknowledged. The warnings were never read. Promberger and Baron (2006) found that patients were more likely to follow medical advice that was provided by a physician than by a computer, and they were less satisfied with the advice if it was provided by a computer. Arkes, Shaffer and Medow (2007) provided evidence that patients think less highly of a physician if the physician consulted a decision aid. “Patients may surmise that a physician who uses a [decision support system] is not as capable as a physician who makes the diagnosis with no assistance from a [decision support system]” (p.189).

Those searching for the causes of decision aid neglect have examined characteristics of the decision maker. As decision makers gain experience, their confidence grows, but that experience does not always translate to improvements in decision making accuracy (Lichtenstein & Fischhoff, 1977; Bradley, 1981; Arkes & Freedman, 1984; Yates, McDaniel & Brown, 1991; Dawson, Connors, Speroff, Kemka, Shaw, & Arkes, 1993). Lack of feedback or biased interpretation of our feedback can give us an inflated opinion of our abilities (Einhorn & Hogarth, 1978; Brehmer, 1980).
Such factors can contribute to our unwillingness to rely on a decision aid (Kleinmuntz, 1990). Increased confidence of the decision maker appears to discourage decision aid reliance. Whitecotton (1996) examined a group of financial experts and financial students and tested their ability to forecast the future success of a company based on its financial profile. Those who expressed more confidence were less likely to rely on the provided decision aid. Arkes, Dawes, and Christensen (1986) asked participants to select which of three baseball players won Major League Baseball’s Most Valuable Player Award in a given year. Four performance statistics were listed for each player. In addition, a decision aid was provided that told participants how to use the statistics to their advantage. Those who knew a great deal about baseball estimated that they answered a high number of the questions correctly. They more often did not use the advice suggested by the aid. Those who knew only a moderate amount about baseball estimated they got a lesser number of the questions correct. They relied on the decision aid more, and because of this, actually outperformed the highly overconfident baseball experts. Sieck and Arkes (2005) asked participants to decide whether a prospective jury member either supported or opposed physician-assisted suicide based on five diagnostic parameters. An equation combined the cues and provided advice to the participant if they chose to consult the advice. They found that participants tended to be highly overconfident in their answers and rarely consulted the equation. If, however, enhanced calibration feedback (showing how overconfident the individual was) was given to the participant, their confidence was effectively lowered and they were found to then consult
the equation more often. The results are consistent. Confident decision makers have faith in their personal judgment, and don’t believe they need the advice of a decision aid.

Further, confident decision makers appear willing to pay a premium to rely on their own judgment versus that of a formula. Heath and Tversky (1991) asked participants a series of general knowledge questions. After each four-part multiple choice question, they were asked to assess how confident they were in their chosen answer: 25% confident for “pure guess” to 100% confident for “absolutely certain.” After stating a confidence level, participants were given a choice: bet that the answer they gave was correct, or bet on an equal chance lottery. In other words, if the participant answered that they were 80% confident that they had given the correct answer, the participant was given the option of a) betting that their answer was correct or b) betting on a lottery in which the chances of winning were 80%. If the person gave a confidence level that was high (about 80% or better) participants overwhelmingly chose to bet on their personal judgment versus the equal chance lottery. They were even willing to pay a premium and chose their personal judgment over a lottery that offered 15%-20% better chances. These results are consistent with the notion that a preference for personal judgment is, at least when confidence is high, the preferred default. If we consider the lottery a stand-in for a decision aid, it suggests that highly confident people will rely on their personal judgment even when it can be shown the decision aid is a bit better. Such a preference was not present when confidence in one’s answer was low however. In fact, participants who were not very confident in their answer tended to choose the equal chance lottery more often than their own personal judgment. In general then (for an
exception, see Arnold, Clark, Collier, Leech, & Sutton, 2004), more confident and more experienced decision makers make less use of decision aids than less confident and less experienced decision makers do.

What about the characteristics of the decision aid? A decision maker’s perception of decision aid validity may also contribute to neglect. Sieck and Arkes (2005) note that “many aids do not produce an output that appears to have an obviously salutary effect on the outcome of the decision” (p. 30). Accuracy of a decision aid over many persons may seem less applicable when the aid is to be applied to an individual. Maybe this individual represents an exception to the rule. Similarly, the aid may request information that appears unnatural to the user. “[T]he decision maker may perceive no apparent link between a new, unusual procedure and a better decision outcome” (p. 30). Regarding decision aids used in the medical field (referred to as Clinical Decision Support Systems or CDSSs), Wendt, Knaup-Gregori, and Winter (2000) note that the validity of a decision support system must be recognizable. “An important problem of the acceptance of CDSSs relates to the fact that the user must be convinced of the validity, i.e. correctness and completeness, of the information provided” (p. 854). Shotland, Alliger, and Sales (1998) observe “We have seen that managers who give selection tests to applicants are themselves uncomfortable with tests which defy everyday logic” (p. 125).

Defining Face Validity

These concepts of whether an aid appears to be having an effect on improving decision making is similar to the concept of face validity in psychometrics. Face validity (FV) is described by Mosier (1947):
The term ‘face validity’ implies that a test which is to be used in a practical situation should, in addition to having pragmatic or statistical validity, appear practical, pertinent and related to the purpose of the test as well; i.e., it should not only be valid, but it should also appear valid (p. 192).

Nevo (1985) describes the measurement of face validity:

A measurement of face validity is made when a rater who is a testee, nonprofessional user, or interested individual (“public”) rates a test item, test, or battery of tests by employing an absolute or relative technique as very suitable (or relevant) to unsuitable (or irrelevant) for its intended use (p. 289).

Anastasi (1988) notes:

[Face validity] is not validity in the technical sense: it refers not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test “looks valid” to the examinees who take it, the administrative personnel who decide on its use, and other technically untrained observers (p. 144).

Nevo (1985) summarizes a basic assertion of many texts that:

FV should be separated from criterion-related, content, or construct validity. FV should not be confused with the other types of validity and it cannot replace them. (p. 287).

It should be acknowledged that the term face validity can still be a source of confusion. Occasionally face validity is defined as a component of content validity (e.g. Haynes, Richard, & Kubany, 1995). Gaber and Gaber (2010) recognize that some
believe face validity should be measured by experts, rather than the untrained (p.139). In this study, the concepts of face validity described by Mosier (1947), Anastasi (1988), and Nevo (1985) will constitute the working definition for the term. Face validity is an assessment made by non-experts. It is an evaluation of the suitability of a test/item to fulfill its stated purpose. It is separate from concepts of construct validity that might be evaluated by test developers or other trained experts.

The Study of Face Validity

In general, test developers have not concerned themselves much with face validity. “Face validity is the Rodney Dangerfield of psychometric variables: It has received little attention—and even less respect—from researchers examining the construct validity of psychological tests and measures” (Bornstein, Rossner, Hill, & Stepanian, 1994, p. 363). Bornstein (1996) notes that while the original 1954 version of the American Psychological Association’s *Technical Recommendations for Psychological Tests and Diagnostic Techniques* included an extensive discussion of face validity\(^1\), the 1985 version makes no mention of face validity at all. Messick’s (1995) influential article proposing a unified model of validity makes no mention of face validity as a component.

Decision aid developers may have more reason to be concerned about the face validity of a decision aid (Table 1). There are several differences between the test taker

\(^1\) The 1954 APA recommendations don’t specifically use the term “face validity” but do discuss test user’s perception of validity.
who is being subjected to a test, and the decision maker who is being subjected to a
decision aid. Test takers don’t typically have a choice on whether they want to take a
particular test. Job applicants, for example, must take the company’s selection tool if
they want a job. This is not the case for most decision aids. The typical decision aid user
can choose to disregard the advice of a decision aid, or not use the decision aid at all.
Another distinction between a test taker and a decision maker is the type of performance
each is attempting to maximize. The test taker is typically motivated to do well on the
test, irrespective of whether he/she believes the test is appropriate. The test taker will
therefore give their due diligence to the test even in instances when they don’t believe the
test is suitable for its stated purpose. On the other hand, the decision maker is motivated
to make a good decision irrespective of what a decision aid might suggest. The decision
maker may readily disregard a decision aid if they do not believe it is suitable for its
stated purpose. In the realm of tests, as long as a measure can be shown to have good
psychometric properties and statistical evidence of construct validity, test developers can
disregard test takers’ perceptions with limited consequence. For those attempting to
courage the use of a decision aid, the issue of face validity may be more important.
Table 1

Test takers versus decision makers

<table>
<thead>
<tr>
<th>Test takers taking a test…</th>
<th>Decision makers using a decision aid…</th>
</tr>
</thead>
<tbody>
<tr>
<td>…must complete the test to get the job/grade/school they want.</td>
<td>…often have the option to not use the aid, or not abide by the advice it gives.</td>
</tr>
<tr>
<td>…are motivated to maximize their test score, and will give the test their full effort even when they don’t believe the test is appropriate.</td>
<td>…are motivated to make the best decision possible, and will readily disregard a decision aid if they don’t believe the aid is appropriate.</td>
</tr>
</tbody>
</table>

Studying decision aid face validity can make a potential contribution to our understanding of decision aid neglect. Some key questions include: 1) do decision makers spontaneously make a judgment about the suitability of a decision aid? 2) Can it be shown that such a judgment of suitability affects the decision maker’s willingness to rely on the decision aid? Chapters two and three will attempt to answer these questions.
Chapter 2: Measuring Decision Aid Face Validity

Assessing the face validity of psychometric measures has been a neglected area of study. The study of face validity of decision aids is also not mature. To measure face validity of psychometric measures, Nevo (1985) suggests an “absolute” approach as well as a “relative” approach. The absolute approach involves assessing a measure on its own merits in isolation. For example, the rater is asked to rate a test or item on a five-point-scale: “5”-the item is extremely suitable for a given purpose; “4”-the item is very suitable for a given purpose; “3”-the item is adequate; “2”-the item is inadequate; and “1”-the item is irrelevant and therefore unsuitable. Secolsky (1987) recommends the addition of a “cannot determine” option in the event that the participant has no basis for an assessment of validity. In the “relative” approach to measuring face validity, a rater simultaneously judges the FV of several tests or items by comparing them to each other and ranking them accordingly.

Study 1: Measuring Face Validity of Decision Aids

The first study adapts Nevo’s scale originally developed to assess the face validity of test items to see if it can be used to assess the face validity of a decision aid as well. I opted to administer the relative approach first to participants, followed by the absolute approach. This order allows participants to view all decision aids at once (during the relative ranking) to better appreciate the range of decision aids they will be exposed to.
Participants can use this knowledge during the later absolute rating procedure, resulting hopefully in reduced respondent variance when making their face validity assessments. I used the data obtained during the later absolute approach for analysis.

 Method-Relative Approach to Assessing Face Validity

Undergraduate psychology students ($n = 156$) took part for course credit. The data from 17 participants were eliminated after they indicated in a post-study survey that “Using my data is probably not a good idea. I answered randomly, or otherwise didn't take the experiment very seriously.” Data from the remaining 139 participants were analyzed. Participants completed the study on personal computers. On the computer screen, participants read the instructions “Assume you want to estimate the monthly rent of a specific apartment in the Columbus, Ohio [local] area. On the left side of the screen are methods you could use to estimate the monthly rent. Please rank the methods from most appropriate for estimating rent to least appropriate for estimating rent by dragging each box from the left side of the screen over to the right hand side of the screen. Place the most appropriate method at the top, and the least appropriate at the bottom.” Participants saw 12 rental estimate methods presented on the left half of the screen in random order. The right half of the screen was empty except for the label “Most appropriate” at the top right half of the screen and “Least appropriate” labeled at the bottom. Participants used their computer mouse to drag each method over to the right hand of the screen in the appropriate rank order. The methods to estimate local rent were:
1) Use a linear regression equation fitted to a sample of 50 random Columbus apartments that uses # bedrooms, sq.ft., # baths.

2) Use the Columbus, Ohio average for apartment rent: Studio: $481; 1-Bed: $540; 2-Bed: $653; 3-bed: $831.

3) Pick the nearest advertised apartment (but not the same complex) with an equivalent number of rooms, and use their advertised rent.


6) Start at $500 for a studio, and add $200 for each additional bedroom.


8) Take 5 random apartments in the same zip code and average them together.

9) Always estimate $800.

10) Draw a number at random out of a hat between $800 and $1500.


Results—Relative Approach to Assessing Face Validity

The mean rank was computed for each method. No method was hypothesized to be more or less face valid than another method a-priori, so statistically significant differences in face validity were not assessed. Ordered from most to least face valid, the methods were ranked (see Figure 1): Columbus Average ($M = 2.04$), Ohio Average ($M = 3.35$), Linear Regression ($M = 3.81$), Nearest Apartment ($M = 4.30$), Cleveland Average ($M = 5.63$), $500 + $200 per room ($M = 6.02$), 5 Random Apartments from same Zip Code ($M = 6.15$), National Average ($M = 6.36$), Always $800 ($M = 9.57$), California Average ($M = 9.76$), Random Number $800-$1500 ($M = 10.46$), Beverly Hills Average ($M = 10.55$). Agreement between the 156 judges was quantified using Kendall’s coefficient of concordance ($W = .64$, $\chi^2(11) = 1100.24$, $p < .001$).
Method-Absolute Approach to Assessing Face Validity

Immediately after completing the relative ranking approach described in the previous section, the same 139 participants were used for the absolute approach to assessing face validity. On a computer screen, participants were told “Please evaluate the suitability of each method individually. How suitable is the given method to estimate the rent of a Columbus Ohio apartment?” Instead of seeing all rental estimate methods at once, participants now saw them one at a time. For each method, participants were asked to rate the method on a Likert scale from 1 to 5 (5: Extremely suitable to estimate rent of a Columbus Ohio apartment; 4: Very suitable to estimate rent of a Columbus Ohio apartment; 3: This method is adequate; 2: This method is inadequate; 1: This method is
irrelevant and therefore unsuitable. Participants could also select a sixth option (NA: Cannot determine). The same 12 rental estimate methods presented in the previous section were again presented in random order.

**Results-Absolute Approach to Assessing Face Validity**

The mean face validity was calculated for each rental estimate method. Listed from most face valid to least face valid (see Figure 2): Columbus Average ($M = 4.34$), Ohio Average ($M = 3.55$), Linear Regression ($M = 3.50$), Nearest Apartment ($M = 3.32$), 5 Random Apartments from same Zip Code ($M = 2.85$), $500 + 200$ per room ($M = 2.74$), Cleveland Average ($M = 2.66$), National Average ($M = 2.59$), Always $800$ ($M = 1.69$), Random Number $800-$1500 ($M = 1.52$), California Average ($M = 1.36$), Beverly Hills Average ($M = 1.30$). Once again, no a-priori hypotheses were made regarding which methods would be judged more face valid, so no statistical comparisons were made. Agreement between judges was similar to that observed during the relative ranking method (Kendall’s $W = .65$, $\chi^2(11) = 896.78$, $p < .001$ with 156 complete ratings).
Discussion

The obtained results seem reasonable. Participants recognize that using the average city rent is more reasonable than using a random number. A comparison of Figure 1 and Figure 2 reveals that the relative rank order of judged face validity of the various methods is very similar no matter if face validity is assessed using the relative ranking method, or the absolute approach (Spearman’s $r_s(10) = .79, p < .001$). The results indicate that Nevo’s scale can potentially be used to measure the face validity of a decision aid.

Figure 2. Study 1. Face validity of various methods to estimate rent of a Columbus, Ohio apartment (absolute rating method). Bars represent 95% confidence intervals.
Chapter 3: Using Face Validity to Predict Decision Aid Neglect

Chapter two examined whether a measure of face validity (FV) developed for tests and test items could potentially be used to evaluate face validity of a decision aid. This chapter examines whether such FV assessments can be used to predict one’s willingness to rely on that decision aid.

Study 2: Face Validity and Decision Aid Neglect

Method

Sample and Procedures

The sample consisted of 129 undergraduate psychology students who participated for course credit. In addition, participants could earn money (up to $12) based on their performance in the task. Data from 12 participants were removed from analysis because the participant had not filled out the survey form according to instructions, or they indicated in a post experiment survey that “Using my data is probably not a good idea. I answered randomly, or otherwise didn't take the experiment very seriously.” This left data from the remaining 117 participants for analysis.

The experiment was conducted on personal computers. On each participant’s screen appeared the instructions: “You will see twelve apartments described to you. Give your best estimate of the total monthly rent of each apartment. To assist you, a series of
decision aids have been developed to estimate apartment rent. Some of these aids are more accurate than others. After your initial estimate of the rent, you will have an opportunity to see one of these decision aids and the estimate it produces. After you see the decision aid and its estimate, you will have an opportunity to revise your initial rent estimate if you wish. A range (plus or minus 10%) will be created around your final guess. If the range includes the actual monthly rent of the apartment, you will earn $1. If the range does NOT include the actual rent, you will lose $1. This process will repeat for each apartment. If you have a positive balance at the end of the experiment, that amount will be given to you in REAL CASH. If you end the experiment with a zero or negative balance, you will not be given any cash. You can potentially earn up to $12 in this experiment.”

Following these instructions, participants saw, presented in random order, information about 12 apartments in the local (Columbus, Ohio) area. For each apartment participants viewed apartment details such as local address, square footage, number of bedrooms and bathrooms, an outside picture of the building, the floor plan, and a description of the local community (see Figure 3 for an example).
Figure 3. Study 2. Example apartment data viewed by participants.

Participants were asked to estimate the monthly rent of the apartment. Next, participants saw the results of a decision aid that also estimated the rent of the same apartment. Participants saw the estimate, as well as how the estimate was derived. For example: “Decision aid predicted rent: $481. The decision aid uses the Columbus, Ohio average for a studio apartment. Your rent estimate was $XXX. Based on this new information, what is your revised estimate? (Enter your original number again if you wish to keep your original estimate unchanged).” In total, there were 12 different decision aids (the same decision aids evaluated in study 1). All decision aids were used to make an estimate for all 12 apartments, resulting in a total of 144 predictions for 12 apartments.
For each apartment, participants saw the result of one (and only one) of the decision aids. Which decision aid seen by the participant was determined randomly. Due to the random selection, participants saw some decision aids multiple times and may not have seen other decision aids at all. Participants did not receive any feedback on the accuracy of their estimates until the end of the experiment, where they were told what their final score was. Those with a positive balance were given cash at the conclusion of the experiment.

The data resulting from this procedure included: 1) the initial rent estimate made by the participant 2) the rental estimate made by the decision aid and 3) the revised rental estimate made by the participant after seeing the decision aid estimate.

Measures.

Reliance on Decision Aid.

The degree to which participants used the advice provided by the decision aid was quantified using the “weight-of-advice” (WOA) measure (Yaniv, 2004; Gino, 2008) where:

$$WOA = \frac{|FinalEstimate - InitialEstimate|}{|Advice - InitialEstimate|}$$

If the judge completely disregards the advice, and keeps the initial estimate unchanged, the resulting WOA equals 0. If the judge heeds the advice completely, and changes their final estimate to match the advice exactly, the resulting WOA is 1. Values between 0 and 1 result when the judge is taking in to consideration both their initial estimate and the advice received. The higher the WOA value, the more relative weight has been given to the advice in making the final decision.
The WOA value cannot be interpreted if the final estimate does not fall between the initial estimate and the advice. For example, if the participant decides to overshoot the advice (ex: Initial=50, Advice=100, Final=150) then a WOA greater than 1.0 results. Likewise, they can move in the opposite direction of the advice (ex: Initial =50, Advice=100, Final=40). In such cases, a WOA value can still be calculated, but the meaning of the value has changed. In addition, if the initial estimate is equivalent to the decision aid advice, then a zero appears in the denominator and reliance on the decision aid cannot be quantified.

Face Validity of Decision Aid.

To measure face validity of the decision aid, data from study 1 was utilized. In study 1, 139 participants evaluated the perceived face validity of 12 decision aids. The obtained scores were used as measures of face validity for the current experiment.

Results

Initial Analyses.

The task of accurately estimating apartment rents within a ±10% range proved to be very challenging. Only two participants earned money ($2 each). In an earlier pilot study, other equally inaccurate participants were on average 59.53% confident that their estimate for a given apartment was within the ±10% range ($n = 139$, $SD = 15.85$). This indicates the task appeared to be achievable, even though the results show otherwise.

The 117 participants produced estimates for all 12 apartments, resulting in 1,404 cases total. In 18 cases, the participant’s initial estimate was the same as the advice provided by the decision aid, and a WOA value could not be calculated. In 126 cases
(9.09% of the remaining cases), the final estimate did not fall between the initial estimate and the decision aid advice. The resulting WOA values from these cases were uninterpretable and deleted from analysis. This left 1,260 cases from 117 participants for analysis. The mean WOA was .28 for the 1,260 apartment estimates.

**Decision Aid Reliance and Decision Aid Face Validity.**

The hypothesis that face validity is correlated to decision aid reliance was supported. Pearson’s $r$ correlation between decision aid face validity (DAFV) and “weight-of-advice” (WOA) was computed for each of the 117 participants. The average Pearson’s correlation for all participants was $r = .23$. A histogram of observed correlations for participants approximates a normal distribution (see Figure 4). To test statistical significance each participant’s correlation underwent Fisher’s $r$ to $z$ transformation (Cohen, 2003). The transformed $z$ scores were averaged across all participants. The mean $z$ of .26 was significantly greater than zero (one sample $t(116) = 7.77, p < .001$).

Since DAFV was measured by an ordinal Likert item, it may be argued that Spearman’s rank order correlation ($r_s$) may be the more appropriate analysis. For each participant, DAFV and WOA were converted to ranks, and Spearman’s $r_s$ was calculated. The average Spearman’s correlation for all participants was $r_s = .20$ (Figure 4). To test statistical significance, each participant’s correlation underwent Fisher’s $r$ to $z$ transformation and the transformed scores were then averaged across participants. The average $z$ of .23 was significantly greater than zero (One sample $t(116) = 6.46, p < .001$). The analysis was repeated using all ($n = 129$) collected data (i.e. no participants were
eliminated for any reason). The statistical conclusions did not change ($r_s = .21$, $z = .23$, $t(128)=7.13$, $p < .001$).

Figure 4. Study 2. Histogram of participant correlations (Decision Aid Face Validity versus Weight-of-Advice).

The data can also be analyzed using a regression approach. The participant’s final estimate was regressed on to the main effects of the initial estimate, the DA advice, and the DA face validity. One advantage of the regression approach is that there is no need to exclude cases where the final estimate does not fall between the initial estimate and the decision aid advice, or if the initial estimate matches the decision aid estimate. Using a
regression approach all 1,404 cases produced by the 117 participants can be included in analysis. All predictor variables were centered using the variable’s grand mean, where:

\[ M_{\text{Init.Est}} = 903.76, M_{\text{DA Adv.}} = 1014.36, \text{ and } M_{\text{DAFV}} = 2.62. \]

The centered terms were used to create one hypothesized interaction term: DAFV * DA Advice. A separate regression was performed for each participant. The resulting regression weights for participants were averaged. The mean regression weight of the hypothesized interaction term (Mean B=0.13, 95% CI: 0.08 to 0.18) was significantly greater than 0 (t(116)=5.47, p<.001). The analysis of all regression terms is detailed in Table 2.

The analysis was repeated using all collected data (n = 129), without eliminating any suspect participants. The interaction term remained significant (Mean B = 0.14, t(128)=5.92, p < .001).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>SE</th>
<th>95% CI</th>
<th>t(116)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>928.44</td>
<td>10.18</td>
<td>[908.28, 948.61]</td>
<td>91.20</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Participant’s Initial Estimate</td>
<td>0.59</td>
<td>0.03</td>
<td>[.53, .65]</td>
<td>19.34</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Advice</td>
<td>0.36</td>
<td>0.04</td>
<td>[.28, .43]</td>
<td>9.48</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Face Validity</td>
<td>42.31</td>
<td>14.14</td>
<td>[14.30, 70.32]</td>
<td>2.99</td>
<td>.003</td>
</tr>
<tr>
<td>DA Adv. * DAFV</td>
<td>0.13</td>
<td>0.02</td>
<td>[.08, .18]</td>
<td>5.47</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

*Note. Study 2. CI = confidence interval.*

Another way to evaluate the significance of the interaction term is to evaluate the change in predictive ability between a model without the interaction term, and another with the interaction term (change in \( R^2 \)). First, a model using only the main effects was
used to predict final rent estimate. The “main effects only” model was derived the same way the “main effect + interaction” model described above was. Final estimate was regressed on to the main effects of Initial Estimate, DA Advice, and DAFV. This regression was done for each participant, and then the regression weights were averaged across all participants. The resulting average regression weights formed the regression equation:

\[
\text{Final Est.} = 902.36 + .62(\text{Init. Est.}) + .28(\text{DA Adv.}) + 6.61(\text{DAFV})
\]

This “main effects only” equation was used to predict final rent estimates. A correlation between the actual final estimate and predicted final estimate was calculated. The correlations were averaged across all participants. The average correlation between the predicted final estimate and the actual final estimate was .77 (\(n=117\), 95% CI: .73 to .81, \(R^2 = .59\)). The “main effects only” equation explained 59% of the variance in the final estimate.

The same procedure was used to make predictions with the equation from Table 2 (main effects + interaction). Where:

\[
\text{Final Est.} = 928.44 + .59(\text{Init. Est.}) + .36(\text{DA Adv.}) + 42.31(\text{DAFV}) + .13(\text{DA Adv.} \times \text{DAFV})
\]

This equation was used to predict a participant’s final rent estimate. The average correlation increased to .79 (\(n=117\), 95% CI: .75 to .83), explaining 62% of the variance in final estimate. In a paired-sample t-test, it was found the equation with the interaction term explained significantly more variance than that explained by the “main effects only” equation. \((t(116)=3.84, p < .001)\).
To visualize the nature of the interaction (Aiken & West, 1991), let’s say the participant makes an initial estimate of $904, which was the mean initial estimate. The obtained regression equation with the interaction term was used to predict the final rent estimate for varying levels of DA advice and DAFV (see Figure 5). When the DA advice is much higher than the participant’s initial estimate, the “job” of the decision aid is to try and convince the participant to raise their initial estimate. The decision aid was able to do this more convincingly when the DA face validity was high. On the opposite end of the spectrum, when the decision aid provided an estimate that was far below that of the participant’s initial estimate, the role of the decision aid is to convince the decision maker to lower their estimate. The table shows that the final estimate is lowered more when the decision aid face validity is high. When the decision aid has low face validity, the decision aid is not as effective in convincing the participant to raise or lower their initial estimate.
Discussion

The participants in the current experiment were not asked to rate the validity of the various decision aids. Yet, the measure of face validity obtained from earlier participants was correlated with the “weight-of-advice” the current participants gave to the aid. This suggests participants in the current experiment spontaneously made their own judgment on the suitability of the various methods, and this judgment affected how much weight they were willing to accord the aid’s advice.
What criteria are participants using to judge face validity? One explanation is they are assessing the reasonableness of the procedure used (i.e. “Using the average of a Columbus, Ohio apartment sounds like a reasonable method”). Another explanation is they are evaluating the actual prediction made by the aid. The participant might think “$2,500 sounds like a really outlandish amount. I think I’ll just go with my original estimate.” Support for this latter view is suggested by literature from the Social Judgment Scheme (SJS) tradition (Davis, 1996). According to SJS, when there is a group of individuals making a joint decision together, the individual tends to lend more weight to others’ opinions that closely match their own opinion. If one individual in the group possesses an opinion that is unique or otherwise unusual in the group, the group tends to discount the opinion. Additionally, Al-Natoor et al. (2008) found that users rate decision aids more highly if the decision aid provides advice that is similar to the users’ initial opinion. In the current experiment the participant may discount the advice provided by the decision aid because the advice provided by the aid is far from the initial opinion formed by the individual. Experiment 3 seeks to control for this possible explanation.

Study 3: DA Procedure as a Determinant of Face Validity and DA Neglect

In study 2, participants had two pieces of information they could use to evaluate the suitableness of the decision aid: the described procedure that the decision aid used and the rent estimate provided by the decision aid. In the current experiment, the rent estimate was made equivalent across all decision aids. The estimate provided by each decision aid was not actually obtained through the described procedure. Rather, the
estimate provided was the actual rent of the apartment. If participants revised their initial estimate to match the estimate provided by the decision aid, they would be 100% accurate. Thus, the primary reason not to follow the decision aid is because the procedure described does not seem appropriate.

Method

Sample and Procedures

The sample consisted of 110 undergraduate psychology students who participated for course credit. In addition, each participant could earn money (up to $20) based on their performance on the experimental task. Participants earned on average $2.51 in the experiment ($SD = $5.39).

The experiment was conducted on personal computers. On the screen, participants read the following instructions: “You will see twenty apartments described to you. Give your best estimate of the total monthly rent of each apartment. To assist you, a series of decision aids have been developed to estimate apartment rent. Some of these aids are more helpful than others. After your initial estimate of the rent, you will have an opportunity to see one of these decision aids and the estimate it produces. After you see the decision aid and its estimate, you will have an opportunity to revise your initial rent estimate if you wish. A range (plus or minus 10%) will be created around your final guess. If the range includes the actual monthly rent of the apartment, you will earn $1. If the range does NOT include the actual rent, you will lose $1. This process will repeat for each apartment. If you have a positive balance at the end of the experiment, that amount will be given to you in REAL CASH. If you end the experiment with a zero or negative
balance, you will not be given any cash. You can potentially earn up to $20 in this experiment.” Participants were then given an opportunity to ask questions.

Participants next saw, presented in random order, information about 20 apartments in the local area. For each apartment participants viewed apartment details such as local address, square footage, number of bedrooms and bathrooms, an outside picture of the building, the floor plan, and a description of the local community. Participants were asked to estimate the monthly rent of the apartment.

After the participant entered an initial estimate of the rent, the participant next saw an estimate made by one (and only one) of eight decision aids. The decision aid methodology was explained, followed by the decision aid rental estimate. For example: “NEAREST SIMILAR APARTMENT: The decision aid references the nearest advertised apartment (but not the same complex) with an equivalent number of rooms, and uses their advertised rent. The NEAREST SIMILAR APARTMENT METHOD predicts this apartment’s rent is $640. Your rent estimate was $XXX. Based on this new information, do you wish to revise your estimate? (enter your original number again if you wish to keep your original estimate unchanged).” In reality, the decision aid always recommended the actual monthly rent of the apartment. Every decision aid made the same (accurate) recommendation for a given apartment. Therefore, the degree to which a

---

2 Four aids from study 2 were not used in study 3. It would have been obvious to participants the aid was not using the method it claimed (e.g., “Always guess $800”).
participant chose to heed the advice was dictated primarily by the degree to which the participant believed the decision aid method was credible.

This process repeated for each of the 20 apartments. Participants did not receive any feedback about the accuracy of their rent estimates until the experiment was over.

Measures

Decision aid face validity

To measure the face validity of a given aid, the current experiment used data from a previous study (study 1) where participants \( n = 139 \) evaluated the perceived face validity of each decision aid used to estimate rent.

Decision aid reliance

To compute decision aid reliance, the previously used “weight-of-advice” (WOA) measure was again utilized (Yaniv, 2004; Gino, 2008). Where:

\[
WOA = \frac{|\text{finalestim} - \text{initialestim}|}{|\text{advice} - \text{initialestim}|}
\]

Results

All 110 participants made estimates for all 20 apartments resulting in 2,200 cases. Time to complete the survey was recorded for each participant. The average time to complete was 9:33 with a range of 4:15 to 16:55. An attempt was made to identify participants who appeared to have completed the survey with unrealistic speed. To normalize the positively skewed distribution of completion times, all times underwent a square-root transformation. Only two participants had a (transformed) time less than two standard deviations below the mean \( (Z = -2.0) \). A more aggressive-than-usual cutoff
score of $Z = -1.0$ was selected in an attempt to exclude participants who likely did not give their full attention to the experiment. Seventeen participants had a transformed Z-score of -1.0 or less. This equates to a time of 6:47 or less. When this experimenter took his own survey, and attempted to mostly read every slide, it took 9:16 to complete. The data from the 17 participants was eliminated from analysis. Additionally, two more participants reported in a post-experiment survey that “Using my data is probably not a good idea. I answered randomly, or otherwise didn’t take the experiment very seriously.” Data from the remaining 91 participants is used in all analyses.

The 91 remaining participants estimated 20 apartments each, for a total of 1,820 cases. In three cases, the initial estimate made by the participant was equal to the decision aid advice, and a WOA could not be calculated. Of the remaining 1,817 cases, there were 174 cases (9.58% of total) where the participant’s final answer did not fall between the initial answer, and the advice provided by the decision aid. In these instances, the WOA statistic is not interpretable. These cases were eliminated from analysis as well, resulting in 1,643 cases (mean WOA = .32) of interpretable data.

For each participant, Pearson’s correlation coefficient was calculated comparing DAFV to WOA. The mean correlation for all participants was $r = .17$. A histogram of the correlations exhibited by participants is provided in Figure 6. The mean correlation (after Fisher’s r to z transformation) was significantly greater than zero (mean $z = .18$, one sample $t(90) = 6.25, p < .001$).

An alternate analysis was also conducted where DAFV was treated as an ordinal variable. WOA and DAFV were converted to ranks for each participant, and Spearman’s
correlation \((r_s)\) was calculated for each participant (see Figure 6). Correlations underwent Fisher’s \(r\) to \(z\) transformation for significance testing. The average correlation \((r_s=.14)\) was significantly greater than zero (mean \(z = .16\), one sample \(t(90) = 5.22, p < .001\)). The analysis was repeated using data from all 110 participants. The conclusions did not change as a result \((r_s=.13, z = .14, t(109) = 4.98, p < .001)\).

Figure 6. Study 3. Histogram of participant correlations (Decision Aid Face Validity versus Weight-of-Advice).

Using a regression approach more of the data can be included in analysis: a total of 1,820 cases from 91 participants. The participant’s final estimate was regressed on to the main effects of the initial estimate, the DA advice, and the DA face validity. These variables were centered using the variable’s grand mean. The centered terms were used to create one hypothesized interaction term: DAFV * DA Advice. A separate regression
was performed for each participant. The regression weights for participants were averaged. The mean regression weight of the hypothesized interaction term (Mean $B=.05$, 95% CI: 0.02 to 0.08) was significantly greater than zero ($t(90)=3.50, p=.001$). The analysis of all the regression terms is detailed in Table 3. The analysis was repeated without eliminating any participants ($n = 110$). The conclusions did not change as a result (Mean $B=.05$, 95% CI: 0.03 to 0.08, $t(109)=4.06, p<.001$).

Table 3

Predictors of Participant Final Rent Estimate

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>SE</th>
<th>95% CI</th>
<th>$t(90)$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>794.35</td>
<td>3.67</td>
<td>[787.06, 801.64]</td>
<td>216.51</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Participant’s Initial Estimate</td>
<td>0.48</td>
<td>0.03</td>
<td>[.43, .53]</td>
<td>17.63</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Estimate</td>
<td>0.53</td>
<td>0.03</td>
<td>[.47, .58]</td>
<td>18.89</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Face Validity</td>
<td>5.69</td>
<td>3.72</td>
<td>[-1.70, 13.08]</td>
<td>1.53</td>
<td>.13</td>
</tr>
<tr>
<td>DA Adv. * DAFV</td>
<td>0.05</td>
<td>0.01</td>
<td>[.02, .08]</td>
<td>3.50</td>
<td>.001</td>
</tr>
</tbody>
</table>

*Note. Study 3. CI = confidence interval.*

As was done in study 2, a regression equation with main effects only was compared to the regression equation that included the interaction term. Final rent was regressed on to the main effects of Initial Estimate, DA Advice, and DAFV. One regression was performed for each participant, and the regression weights were averaged across participants. The resulting equation is given by:

$$\text{FinEst} = 793.73 + .47(\text{InitEst}) + .53(\text{DA Est}) + 3.86(\text{DA FV})$$

Predictions calculated using this formula were correlated with the actual final estimates. One correlation (and one $R^2$) was calculated for each participant, and then all correlations...
(and $R^2$ values) were averaged across participants. The average correlation between the predicted final and the actual final was very high ($r = .91690$). The three main effects, without the benefit of the interaction term, were able to explain 85% of the variance in final rent estimate (average $R^2 = .8474$). This is a tough standard for the interaction term to improve upon. With the interaction term added, the predicted values and the actual values are correlated .91875 (average $R^2 = .8510$). An additional .36% is explained by the addition of the interaction term. This small amount still represents a statistically significant increase in average proportion of variance ($R^2$) explained (Paired Sample $t(90)=2.076, p<.05$). Graphing the interaction shows a replication of the relationship observed in experiment 2 (see Figure 7). For those times that the decision aid estimate is higher than the participant’s initial answer, the DA is more convincing in getting the participant to revise their initial estimate upward if the DAFV is higher. The opposite is true when the DA provides an estimate that is lower than the participant’s initial answer. The DAs with higher FV are more convincing in getting the participant to revise their initial estimate downward.
Figure 7. Study 3. Predicted final rent estimate: Interaction plot (Decision Aid Advice * Decision Aid Face Validity) where: Final Est. = 794.35 + .48(Init. Est.) + .53(DA Adv.) + 5.69(DAFV) + .05(DA Adv. * DAFV). Input variables are centered before inclusion. M_{Init.Est.}=795.82. M_{DA Adv.}=785.50. M_{DAFV}=2.76

Discussion

The results are consistent with the notion that decision aid procedure is at least one significant determinate of decision aid face validity and decision aid reliance. The results of study 2 were replicated in the current experiment. Again, decision aid face validity, as assessed by one group of participants, was a significant predictor of decision aid use in a second group of participants.
Since the current experiment equated the advice provided by all decision aids (i.e., the rent estimate), the description of the procedure used by the decision aid is the most important variable remaining to explain differences in decision aid use. The results reaffirm the idea that participants spontaneously evaluate the suitability of a decision aid procedure, and this evaluation is positively related to one’s reliance on the decision aid. Participants pay attention to the decision aid procedure, and evaluate its suitability. This assessment then affects their willingness to rely on the aid.

The strength of the interaction term (DAFV * DA Estimate) degraded between study 2 and study 3. In study 2, the average regression weight was .12 (95% CI .08 to .18). In study 3 the average regression weight dropped to .05 (95% CI .02 to .08). As stated earlier, participants may gauge the suitability of a decision aid based on the description of the procedure the decision aid uses, or by the reasonableness of the amount the decision aid has suggested. In study 3, all decision aids suggested the same amount. Since the change in experimental procedure corresponded with a drop in the magnitude of the interaction term, the results suggest participants use both of these sources of information (procedure and amount) to gauge the suitability of the decision aid.

The previous experiments provide evidence that participants use the decision aid procedure to judge face validity of the decision aid, and this evaluation helps determine their reliance on the decision aid. Going beyond this simple correlation, can the study of decision aid face validity provide additional theoretical insight? These next studies hope to explore that question.
As was discussed in chapter 1, a well studied variable that affects decision aid reliance is confidence of the decision maker (Arkes, Dawes & Christensen, 1986; Heath & Tversky, 1991; Whitecotton, 1996; Sieck & Arkes, 2005). In general, the more confident one is in their knowledge of the domain of interest, the less likely one is to rely on the input of a decision aid. This general finding can potentially be fine-tuned with the additional variable of face validity of the decision aid. It is hypothesized that when confidence is high, decision aid reliance will be low. When confidence is low, decision aid reliance will be high when decision aid face validity is high, and reliance will be low when decision aid face validity is low (see Figure 8). The proposed interaction effect is tested in the following two studies.
Study 4: Obtaining Measures of Confidence and Decision Aid Face Validity

The objective of study 4 is to obtain ratings of face validity and confidence for stimuli that will be used in study 5.

Method

Undergraduate psychology students \((n = 128)\) participated for course credit. Data from 13 participants were removed from analysis because in a post-study survey they indicated “Using my data is probably not a good idea. I answered randomly, or otherwise didn't take the experiment very seriously.” Data from the remaining 115 is used for all analysis. The study was administered on computers.
Participants were told “Imagine that you have been shown a list of apartments along with information about each apartment. You have been asked to estimate the rent of the apartments. The apartments are of varying sizes and number of rooms. The apartments are located in DIFFERENT CITIES. On the following slides, rate your ability to estimate the rent of apartments located in a particular city.” Participants viewed one example of sample apartment data which included square footage, floor plan, outside picture, and city.

Participants were asked to rate their ability to estimate apartment rents in each of four cities. The four cities were presented in random order and included Columbus, Ohio; Cleveland, Ohio; Beverly Hills, California; and London, England. It was presumed that participants would be more confident about prices in Columbus, Ohio (where the study took place), and less confident about rent prices in Beverly Hills and London. Each city was evaluated by a Likert item from ‘1’ “My rent estimates would be nowhere near the actual rent. I have no sense of what rent might cost in <City X>.” to a ‘5’ “My rent estimates would be fairly accurate. I have a good sense of what rent might cost in <City X>.”

Using the same methodology as was used in study 1, participants were next asked to evaluate the suitability of 12 methods to estimate rent in each of the four cities. A total of 4 face validity scores were recorded for each method (one score for each city).

Results

Four cities were chosen to try to elicit high confidence rent estimates, moderate confidence rent estimates, and low confidence rent estimates. The results indicate the
manipulation had the intended effect (see Figure 9). On a five point measure of confidence, (where ‘5’ represents high confidence in one’s estimates) participants reported an average of 3.76 ($SD = 1.07$) that their rent estimates for Columbus, Ohio (where the participants lived) would be relatively accurate. Cleveland, Ohio was next at 2.89 ($SD = 1.20$), followed by Beverly Hills ($M = 2.40, SD = 1.11$) and London ($M = 1.79, SD = 1.02$). The four cities produced significantly different confidence ratings to estimate rent ($F(3,456)=65.29, p < .001$).

Figure 9. Study 4. Participants’ reported confidence in estimating rent in various cities. Error bars represent 95% confidence intervals.
A total of 12 methods that could be used to estimate the rent of an apartment were described to participants. Four mean face validity scores (one for each city) were computed for each rent estimate method. The face validity scores are displayed graphically in Figure 10.

*Figure 10. Study 4. Face validity of methods to estimate rent in four cities (absolute rating method). Error bars represent 95% confidence intervals.*
Discussion

This study served as a pilot study to prepare stimuli for study 5. The stimuli succeeded in identifying locations that would elicit high, moderate, and low confidence rent estimates by participants. Participants recognized that some methods were more appropriate in some locations than others. Examining Figure 10, one can see participants perceived that it made sense to use the Columbus average when estimating the rent of a Columbus apartment, but this didn’t seem suitable when estimating a London apartment. More importantly for the next experiment, the current study identifies three methods that are evaluated as equally highly suitable, moderately suitable, and not suitable for all cities, regardless of city location (see Figure 11). Using the nearest approximate apartment as a guide was seen as equally suitable ($F(3,455) = .08, p = ns$) in all cities. Randomly picking 5 apartments in the same zip code ($F(3,452) = .30, p = ns$) and picking a random number out of a hat ($F(3,428) = .16, p = ns$) were also not viewed to differ in appropriateness according to city. The face validity scores for these three methods were collapsed across cities and would be used to illustrate a highly face valid, moderately face valid, and low face valid rent estimate method for study 5.
Figure 11. Study 4. Three methods for estimating rent whose face validity does not vary by city. Error bars are 95% confidence intervals.

Study 5: Confidence, Face Validity, and Decision Aid Neglect

Using the stimuli tested and measures obtained from the previous study, we are now able to test the proposed interaction between confidence and decision aid face validity. It is proposed that decision aid use will be uniformly low when decision aid face validity is low. When face validity is high, decision aid use will increase as confidence decreases (see Figure 8).
Method

Sample and procedure

The sample consisted of 119 undergraduate psychology students. The participants took part for course credit, and also had the opportunity to win $10. At the end of each data collection session, all participants who finished the task with a positive score entered a lottery. The number of persons in each session varied between 1 and 12. The winner of the lottery (the person who drew the largest number out of a can) received the $10 prize.

The experiment and all instructions were administered via personal computer. Participants were instructed “You will see twenty apartments described to you. Give your best estimate of the total monthly rent of each apartment. The apartments are in 4 different cities: Columbus, Ohio; Cleveland, Ohio; Beverly Hills, California; and London, England. To assist you, a series of decision aids have been developed to estimate apartment rent. Some of these aids are more helpful than others. After your initial estimate of the rent, you will have an opportunity to see one of these decision aids and the estimate it produces. After you see the decision aid and its estimate, you will have an opportunity to revise your initial rent estimate if you wish. A range (plus or minus 10%) will be created around your final guess. If the range includes the actual monthly rent of the apartment, you will earn 1 point. If the range does NOT include the actual rent, you will lose 1 point. This process will repeat for each apartment. If you have a positive score at the end of the experiment, you will enter a lottery with everyone else who has a positive balance in this experimental session. One person in each session
will win $10 in REAL CASH. If you end the experiment with a zero or negative balance, you will not be eligible for the lottery.”

Participants next saw details of an apartment (selected at random) from one of the four cites and asked to estimate the rent of the apartment. Next, participants viewed one of three decision aids selected at random. The decision aid method was described to participants, and the estimate provided by the aid was displayed. Participants then had the opportunity to revise their initial estimate if they wished. In reality, all decision aids provided the actual rent of the apartment. The three decision aids included:

1. NEAREST SIMILAR APARTMENT: The decision aid references the nearest advertised apartment (but not the same complex) with an equivalent number of rooms, and uses their advertised rent.

2. RANDOM APARTMENTS IN SAME ZIP CODE: This decision aid takes 5 random apartments in the same zip code and averages them together.

3. RANDOM NUMBER: The decision aid draws a number at random out of a hat.

Measures

Face validity of decision aid

The ‘NEAREST SIMILAR’ method was pre-tested in study 4 to have high face validity (M = 3.95 out of 5) that did not differ by city. The method ‘SAME ZIP’ also did not differ by city, and had moderate face validity (M = 3.34). The ‘RANDOM NUMBER’ method was pre-tested to have poor face validity (M = 1.33) that also did not
differ by city. These earlier obtained values will be used to predict decision aid reliance in the current experiment.

Confidence in ability to estimate rent.

Participants (as tested in study 4) were highly confident in their ability to estimate rent in Columbus, Ohio ($M = 3.76$ on 5 point scale). This was followed by Cleveland ($M = 2.89$), and then Beverly Hills ($M = 2.40$). They were least confident in their ability to estimate rent in London ($M = 1.79$).

Reliance on decision aid.

“Weight-of-advice” (WOA) was used as a measure of decision aid reliance.

Where: $WOA = \frac{|\text{finalestimate} - \text{initialestimate}|}{|\text{advice} - \text{initialestimate}|}$

Results-Initial analysis

The sample started with data from 119 participants. The same procedure in previous studies was used to identify participants who had completed the survey with unrealistic speed. The time to complete the survey was recorded for each participant. On average, the survey took 11:23 to complete ($SD = 3:02$). The times underwent a square-root transformation to normalize the distribution of scores. Only two participants had a (transformed) time that was less than two standard deviations below the mean ($Z < -2.0$). Therefore, a more aggressive cutoff score ($Z < -1.0$) was selected. This equates to a time of 8:16 or less. Seventeen participants met this criteria and were eliminated from further analysis. Data from an additional 10 participants were removed because they indicated in a post-experiment survey that “Using my data is probably not a good idea. I didn't
understand what I was supposed to do, I answered randomly, or otherwise didn't take the experiment very seriously.” This left data from 92 participants for analysis. The 92 participants made estimates for 20 apartments each, resulting in 1,840 cases. A computer programming error when the stimulus was created (the wrong value for DA advice was presented) resulted in the loss of data for one question (out of the 20 questions total) for 13 participants. The data from the other 19 questions were analyzed for these participants, leaving 1,827 cases from 92 participants for analysis.

Results

Decision Aid Face Validity and Decision Aid Reliance

The relationship between DAFV and WOA was investigated using the same procedure used in studies 2 and 3. DAFV measures were obtained in study 4 and WOA values were obtained with new participants in study 5. The 92 participants from study 5 produced 1,827 valid estimates for apartments. In 6 cases, the initial estimate matched the DA advice, and a WOA value could not be calculated. Of the remaining 1,821 cases, the final estimate did not fall between the initial estimate and the DA advice in 158 cases (8.68%), confusing the interpretation of the WOA value. These cases were dropped, leaving 1,663 interpretable cases from 92 participants. For these 1,663 cases, the mean WOA = .42.

Pearson’s $r$ and Spearman’s $r_s$ were calculated for each participant for the two variables of interest: DAFV versus WOA (see Figure 12). The average Pearson’s correlation ($r = .25$) was significantly greater than zero, tested after undergoing Fisher’s $r$ to $z$ transformation ($z = .29$, one sample $t(91)=7.68, p < .001$). Spearman’s rank order
correlation (Mean $r_s = .24$), was also significantly above zero ($z = .27$, one sample $t(91)=7.06, p<.001$).

The data were also again analyzed duplicating the regression approach from study 2 and 3. All 1,827 valid cases produced by the 92 participants were included in the regression analysis. The participant’s final estimate was regressed onto the main effects of initial estimate, DA advice, DAFV, and the interaction term of DA Advice * DAFV. A regression was performed for each participant. Once again, the average regression

*Figure 12. Study 5. Histogram of participant correlations (Decision Aid Face Validity versus Weight-of-Advice).*
weight for the interaction term was significant (Table 4). Higher face validities were associated with larger reliance on the decision aid advice (Figure 13).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>SE</th>
<th>95% CI</th>
<th>n(91)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1948.36</td>
<td>13.42</td>
<td>[1921.69, 1975.03]</td>
<td>145.13</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Participant’s Initial Estimate</td>
<td>.39</td>
<td>0.03</td>
<td>[.33, .45]</td>
<td>13.00</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Estimate</td>
<td>.65</td>
<td>0.03</td>
<td>[.60, .70]</td>
<td>24.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Face Validity</td>
<td>34.85</td>
<td>12.65</td>
<td>[9.72, 59.99]</td>
<td>2.76</td>
<td>.007</td>
</tr>
<tr>
<td>DA Adv. * DAFV</td>
<td>.06</td>
<td>0.01</td>
<td>[.04, .08]</td>
<td>5.98</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

*Note.* Study 5. CI = confidence interval.
Figure 13. Study 5. Predicted final rent estimate: Interaction plot (Decision Aid Advice * Decision Aid Face Validity) where: Final Est. = 1948.36 + .39(Init. Est.) + .65(DA Adv.) + 34.85(DAFV) + .06(DA Adv. * DAFV). Input variables are centered before inclusion. M_{Init.Est.}=1802.08. M_{DA Adv.}=1982.77. M_{DAFV}=2.84.

Discussion- Decision Aid Face Validity and Decision Aid Reliance

Decision aid face validity was again observed to be positively correlated to decision aid reliance. Face validity assessments made by one group of participants was significantly correlated to the reliance another group of participants placed on a decision aid. This is the third replication of the effect. Studies 2, 3 and now 5 all resulted in similar findings. Higher face validity is associated with higher decision aid reliance.
Results-Confidence and Decision Aid Reliance

Data included 1,663 interpretable WOA values (i.e. the final estimate fell between the initial estimate and the decision aid estimate) from 92 participants. A graph of the mean WOA (obtained from study 5) versus confidence level (obtained from study 4) indicates a general decrease in decision aid reliance as confidence increases (see Figure 14).

Figure 14. Study 5. Mean “weight-of-advice” for various confidence levels. Error bars represent 95% confidence intervals.
To evaluate the statistical significance of the relationship between confidence and WOA, the same procedure was applied that was used to investigate the relationship between DAFV and WOA. Correlations were calculated for each participant (see Figure 15). Correlations then underwent Fisher’s $r$ to $z$ transformation, and the transformed scores were then averaged. The Pearson correlation (mean $r = -.10$) was statistically less than zero ($z = -.10, t(91)=-3.60, p<.001$). When scores were converted to ranks, and Spearman’s $r_s$ was calculated, the result (average $r_s = -.11$) was also statistically less than zero ($z = -.12$, one sample $t(91) = -3.98, p<.001$).

![Figure 15. Study 5. Histogram of participant correlations. (Confidence versus Weight-of-Advice).](image)
Analyzed another way, four means were computed for each participant: the mean ‘weight-of-advice’ when confidence was low (London), moderately low (Beverly Hills), moderately high (Cleveland) and high (hometown: Columbus). In a repeated measures ANOVA, the mean WOA varied significantly across level of confidence ($F(3, 273) = 5.25, p = .002$).

The data were analyzed again using the regression approach. The regression approach used all 1,827 valid cases produced by the 92 participants. The participant’s final estimate was predicted using the participant’s initial estimate, the confidence level (for a given city as rated by other participants), and the decision aid rent estimate. In addition, the hypothesized interaction term (Confidence * DA Advice) was also included. The average regression weight for the interaction term ($Mean B = -.002$) was not statistically significant ($t(91) = -.10, p=.92$, see Table 5). The effect of confidence was minimal (see Figure 16). Repeating the analysis using only the 1,663 cases where the final estimate fell between the initial estimate and the DA advice (as was done during the correlation analysis) did not change the significance of the interaction term ($Mean B = -.03, SE = -.03, 95\% CI: -.08 to .01$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>SE</th>
<th>95% CI</th>
<th>t(91)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1938.79</td>
<td>23.35</td>
<td>[1892.41, 1985.17]</td>
<td>83.03</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Participant’s Initial Estimate</td>
<td>.40</td>
<td>0.03</td>
<td>[.34, .46]</td>
<td>13.61</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Advice</td>
<td>.62</td>
<td>0.03</td>
<td>[.56, .69]</td>
<td>18.60</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Confidence</td>
<td>-8.53</td>
<td>31.54</td>
<td>[-71.18, 54.11]</td>
<td>-.27</td>
<td>.79</td>
</tr>
<tr>
<td>DA Adv. * Confidence</td>
<td>.00</td>
<td>0.03</td>
<td>[-.06, .05]</td>
<td>-.10</td>
<td>.92</td>
</tr>
</tbody>
</table>

*Note.* Study 5. CI = confidence interval.
Figure 16. Study 5. Predicted final rent estimate: Interaction plot (Decision Aid Advice * Confidence) where:

Discussion—Confidence and Decision Aid Reliance

As reviewed in chapter 1, past research has found a negative relationship between decision maker confidence, and willingness to rely on a decision aid. The current results replicate these findings, although one method of analyzing the data resulted in an non-
significant finding. In the current data, confidence was found to be significantly negatively correlated with weight-of-advice, according to Pearsons’ $r$, Spearman’s $r_s$, and ANOVA results. However, when final rent estimate was regressed on to the main effects and the interaction term of interest:

$$Final\ Rent\ Estimate = b_0 + b_1(Init\ Est)+b_2(DA\ Adv.)+b_3(Conf)+b_4(DA\ Adv\ *\ Conf)$$

the interaction term was not significant (Table 5). I offer two interpretations of why confidence might be a non-significant term in the regression equation, but not the other analyses.

The first potential factor contributing to a non-significant result in the regression analysis is the absence of the hypothesized three-way interaction. It was hypothesized that confidence would play a role in decision aid reliance when DA face validity was high, but not when DA face validity was low. This distinction is not made in the current regression analysis. The significant result in the ‘high face validity’ condition may be washed out by the non-significant ‘low face validity’ condition. This limitation should also apply to the correlation and ANOVA analyses as well. The explanation does not therefore explain why the correlation/ANOVA results are significant and the regression result is not.

The second possible explanation relates to the diffusion of an effect as we move down the causal chain. To illustrate with an analogous example, Garg et al. (2005) conducted a review examining the effects of medical decision support systems on physician performance and patient outcomes. The researchers found that decision support systems did indeed improve physician performance in 64% of the 97 studies
examined. However, improvement in patient outcomes was only observable in 13% of the studies examined. Why the drop in observable effect? The presence of a decision support system is only one factor contributing to physician performance. Physician performance, in turn, is only one factor contributing to patient outcome. Thus, the impact that a decision support system has on the proximal variable of physician performance is easier to observe than the effect of a decision support system on the more distal variable of patient outcome. As additional links are added to the causal chain, the effect of the manipulation is further diffused with other factors. The additional link in the causal chain may make it more difficult to see the hypothesized relationship. In any event, the hypothesized three-way interaction between confidence, DA face validity, and DA reliance is tested more directly in the next analysis.

*Results-Decision Aid Face Validity, Confidence, and Decision Aid Reliance*

When using DAFV and confidence (as assessed by previous participants) to predict decision aid reliance, it was predicted that an interaction would be observed such that confidence would be negatively correlated to WOA when DAFV was high. When DAFV was low, WOA would be always low-no matter the level of confidence (see Figure 8). Looking at the obtained results, the proposed interaction is suggested when the mean WOA values are plotted (see Figure 17).

Recall that the “weight-of-advice” measure is not interpretable if the final estimate does not fall between the initial estimate and the decision aid advice. This distinction was ignored in Figure 17. All obtained WOA values, both interpretable and

57
not interpretable, were used to calculate the sample means. In Figure 18, the uninterpretable WOA values were removed before the sample means were calculated. Another way to deal with uninterpretable WOA values, instead of deleting them, is to assign them a value of ‘0’ or ‘1’ as appropriate. If the participant chooses to revise their estimate in the opposite direction, away from the estimate provided by the decision aid, a WOA value of ‘0’ can be assigned. If the participant decides to ‘overshoot’ the advice, a WOA value of ‘1’ can be assigned. In this way, the new WOA value provides a better representation of whether the decision aid was relied upon or not, as opposed to automatically using the value obtained using the WOA equation. In Figure 19, this change has been made to uninterpretable WOA values, and the sample means calculated. The three figures help to illustrate that appropriately modifying or deleting improper WOA values does not change the basic nature of the relationship of the variables.
Figure 17. Study 5. Plot of mean “weight-of-advice” measures obtained from participants for varying levels of confidence and decision aid face validity. All obtained WOA values were used to compute mean statistics. Error bars represent 95% confidence intervals.

Figure 18. Study 5. Plot of mean “weight-of-advice” measures obtained from participants for varying levels of confidence and decision aid face validity. Only valid WOA values were used to compute mean statistics, where the final estimate was between the initial estimate and the decision aid advice. Error bars represent 95% confidence intervals.
A regression approach was used to evaluate the statistical significance of the interaction. Weight-of-advice was regressed on to (centered) DAFV, (centered) Confidence, and the interaction term DAFV * Confidence. As has been done in previous studies in this paper, uninterpretable WOA values (8.68% of cases) were deleted from the analysis. The remaining 1,663 cases from 92 participants were used to evaluate the proposed ‘DAFV’ by ‘Confidence’ interaction. One regression equation was created for each participant based on that participant’s responses. The regression weights were then averaged across all participants (see Table 6). The regression weight of the interaction term was significantly less than zero ($t(91) = -3.03, p = .003$). To illustrate the nature of the interaction, the equation was used to predict WOA values for various levels of
decision aid face validity and confidence (see Figure 20). The graph supports the hypothesized interaction (Figure 8). When DAFV is low, DA reliance is flat and low regardless of confidence level. When DAFV is high, DA reliance is also high when DM confidence is low, but DA reliance gradually decreases as DM confidence increases. The analysis was repeated with all participants’ data included. The results did not change. The interaction term remained significant (Mean B = -.03, t(118) = -3.37, p = .001).

Table 6

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>SE</th>
<th>95% CI</th>
<th>t(91)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>.43</td>
<td>.01</td>
<td>[.40, .46]</td>
<td>26.06</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Face Validity</td>
<td>.07</td>
<td>.01</td>
<td>[.05, .09]</td>
<td>7.40</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Participant Confidence</td>
<td>-.05</td>
<td>.01</td>
<td>[-.07, -.02]</td>
<td>-3.90</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DAFV * Confidence</td>
<td>-.03</td>
<td>.01</td>
<td>[-.06, -.01]</td>
<td>-3.03</td>
<td>.003</td>
</tr>
</tbody>
</table>

Note. Study 5. CI = confidence interval.
Figure 20. Study 5. Predicted decision aid reliance for various levels of confidence and decision aid face validity. Where: Predicted WOA = 0.43 + 0.07(DAFV) − 0.05(Conf.) − 0.03(DAFV * Conf.). Input variables are centered before inclusion. M_{DAFV}=2.82. M_{Conf}=3.76. WOA = “weight-of-advice” where: 

\[ WOA = \frac{\text{FinalEstimate} - \text{InitialEstimate}}{\text{Advice} - \text{InitialEstimate}} \]

To mirror the regression analysis used in previous studies in this paper (where the dependent variable was “final rent estimate”), a second regression analysis was completed. Participant’s final rent estimate was regressed on to four variables (participant’s initial estimate, decision aid advice, decision aid face validity, and confidence), three 2-way interactions (DAFV * DA Advice; Confidence * DA Advice; DAFV * Confidence) and one three-way interaction (DAFV * Confidence * DA Advice). The three-way interaction is the term of interest for testing the hypothesized relationship in Figure 8.
Since WOA measures are not used, this regression approach does not need to be concerned about inappropriate WOA values. All 1,827 cases were analyzed from 92 participants. One regression equation was created for each participant. The regression weights were averaged across all participants. The average regression weight of the hypothesized 3-way interaction (\( \text{Mean } B = -.08, \text{ SE } = .03 \)) was significantly less than zero (\( t(91) = -2.30, p < .05 \)). The analysis of all regression weights is included in Table 7.

### Table 7

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>SE</th>
<th>95% CI</th>
<th>( t(91) )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1973.57</td>
<td>29.83</td>
<td>[1914, 2032.83]</td>
<td>66.15</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Participant’s Initial Estimate</td>
<td>0.38</td>
<td>0.03</td>
<td>[.32, .44]</td>
<td>13.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Decision Aid Advice</td>
<td>0.68</td>
<td>0.03</td>
<td>[.61, .74]</td>
<td>20.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DA Face Validity</td>
<td>-15.57</td>
<td>29.32</td>
<td>[-73.81, 42.67]</td>
<td>-0.53</td>
<td>.60</td>
</tr>
<tr>
<td>Confidence</td>
<td>67.04</td>
<td>39.92</td>
<td>[-12.26, 146.35]</td>
<td>1.68</td>
<td>.10</td>
</tr>
<tr>
<td>DA Adv. * DAFV</td>
<td>.03</td>
<td>0.03</td>
<td>[-.02, .08]</td>
<td>1.31</td>
<td>.20</td>
</tr>
<tr>
<td>DA Adv. * Conf</td>
<td>.05</td>
<td>0.03</td>
<td>[-.01, .12]</td>
<td>1.64</td>
<td>.10</td>
</tr>
<tr>
<td>Conf * DAFV</td>
<td>-47.65</td>
<td>44.09</td>
<td>[-135.23, 39.93]</td>
<td>-1.08</td>
<td>.28</td>
</tr>
<tr>
<td>DA Adv. * DAFV * Conf</td>
<td>-.08</td>
<td>0.03</td>
<td>[-.15, -.01]</td>
<td>-2.30</td>
<td>.02</td>
</tr>
</tbody>
</table>

**Note.** Study 5. CI = confidence interval.

The three way interaction is graphed in Figure 21. When confidence is low, face validity appears to play a role in how much the participant will rely on the aid. Decision aids that are highly face valid are more effective in getting participants to raise or lower their initial estimate. When confidence is high, the importance of face validity fades, and participants no longer strongly differentiate between highly face valid aids and less face valid aids.
Figure 21. Study 5. Predicted final rent estimate as a function of DA Advice for varying levels of confidence and DA face validities. Hypothetical initial estimate of $1,802. Final Est = 1973.57+.38(Init Est)+.68(DA Adv)-15.57(DAFV)+67.04(Conf)+.03(DA Adv*DAFV)+.05(DA Adv*Conf)-47.65(Conf*DAFV)-.08(DA Adv*DAFV*Conf). Predictor variables are centered before inclusion. MInit = 1802.08. MDA Adv = 1982.77. MDAFV = 2.84.

Discussion—Decision Aid Face Validity, Confidence, and Decision Aid Reliance

The data supports the hypothesized interaction. When confidence is low, participants will rely on a face valid decision aid, but they are less willing to rely on a decision aid that is not face valid. When confidence is high, reliance on a decision aid is
relatively low and unaffected by decision aid face validity. This relationship is especially
well illustrated in Figure 20, which compares very favorably to the hypothesized
relationship in Figure 8.
Chapter 5: Influencing Decision Aid Face Validity

One way decision makers (DMs) may infer validity is how the performance of the decision aid (DA) is described. Performance can be communicated in a variety of ways, and some forms may imply greater validity than others. For instance, saying that a given aid is “70% accurate” (to be referred to as the ‘DA only’ condition) is likely a difficult benchmark for the decision maker to judge. The decision maker asks “Is this better than I can do, or worse?”

Even when performance of a decision aid is compared to the typical performance of unaided decision makers, the form of that communication can still present a variety of messages. For instance, one can communicate the accuracy in terms of the proportion of decision makers beaten: “This decision aid performs better than 70% of all participants” implies that 30% do better than the decision aid. The “better than average” effect (Larwood & Whittaker, 1977; Svenson, 1981; Weinstein, 1987) suggests many individuals may assume they are in the top 30% of decision makers. One might also additionally provide the mean performance of DMs for comparison (for short, ‘DA&DM’ condition): “The decision aid was, on average, 70% accurate compared to 60% accuracy of unaided decision makers.” This mode of communicating performance indicates the decision aid is better than typical unaided decision makers, but masks the possibility that some decision makers do better than the aid.
One single communication mode likely doesn’t imply the highest validity in all situations. Imagine the normal distribution of performance that is centered on the mean performance of unaided decision makers. If the variance of this distribution is very small, all decision makers perform about the same and even the most proficient decision makers are not much better than average. In such a tight distribution, very few decision makers may perform well enough to beat the decision aid. In such an event, communicating the proportion of DMs beaten by the aid would be advantageous to the perceived decision aid validity. For instance: “The decision aid performed better than 95% of unaided decision makers.” If decision maker performance has a large variance, the best decision makers are much better than the mean and proficient enough to outperform the aid. This change occurs even though mean performance of decision makers remains unchanged (see Figure 22). Under this condition, the proportion of DMs beaten by the aid seems less impressive: “The decision aid performed better than 60% of unaided decision makers.” In this latter case, higher decision aid face validity may be implied by instead describing the mean performance of DA and DM: “The decision aid had an overall accuracy of 80%. Unaided decision makers had an overall accuracy of 70%.”
All of this assumes decision makers actually attend to such decision aid performance information when assessing the suitability of a given decision aid, a premise that will be tested in the next study.
Study 6: Communicating DA Performance and its Effect on DA Face Validity

Several a-priori hypotheses were made. First: presenting the typical DA performance to participants as well as the typical DM performance (‘DA&DM’ condition) will generally lead to a higher level of assessed face validity versus communicating only the mean DA performance alone. Given no benchmark for how well unaided decision makers perform (as in the ‘DA only’ condition), it is predicted that participants will generally overestimate how well decision makers would do in such a task and thus view the decision aid as less useful when only DA performance is presented.

It is also hypothesized that communicating the proportion of DMs beaten by a DA (‘% DMs beaten’ condition) will vary in persuasiveness depending on the variance of DM performance. For a given mean performance, when DM variance is high, relative proportions will be less convincing than when variance is low (see Figure 22).

Lastly, there is a potential interaction such that the disparity in face validity (between ‘DA&DM’ and ‘DA only’ conditions) is much higher when the overall accuracy of the decision aid is low (see Figure 23). For instance, imagine that you hear a particular decision aid is accurate only 60% of the time. This sounds horrible until you are told that unaided decision makers are only 50% accurate. Given this additional information, your opinion of the decision aid improves. Alternatively if you are told that a given decision aid is accurate 95% of the time, this sounds pretty good to you. If additional information is provided that unaided decision makers are accurate 85% of the time, this will likely not much affect your positive opinion of the decision aid.
Method

Two-hundred and two undergraduate psychology students participated for course credit. The experiment was administered via personal computers. Data from forty-five participants was removed from analysis because, in a post-experiment survey, they indicated “Using my data is probably not a good idea. I didn't understand what I was supposed to do, I answered randomly, or otherwise didn't take the experiment very seriously.” Data from two additional participants were eliminated because the time spent on the survey was less than 2 standard deviations below the mean time spent on the
survey (computed after all times underwent a square root transformation to normalize the distribution of scores). The eliminated participants spent 2:51 and 3:21 on the survey, compared to a mean completion time of 7:31. Data from the remaining 155 participants is used in all analysis.

A series of hypothetical decision aids was presented to participants (see Appendix A). The performance of the aid was communicated in one of three ways: proportion of DMs beaten by the DA (“This decision aid performed better than 76% of human resource experts who did not use the aid.”), mean DA performance (“The decision aid had an overall accuracy of 70%”), or mean ‘DA&DM’ performance (“The decision aid had an overall accuracy of 70%. Unaided admissions officers had an overall accuracy of 60%.”). The “Proportion beaten” condition had two versions: one for theoretical high DM performance variance, and another for low performance variance. Changing the variance (high or low) does not change the DA or DM mean performance.

There were a total of 5 decision aids. Each participant saw the performance of a single aid described in four different ways: (proportions-high variance, proportions-low variance, ‘DA only’ mean and ‘DA&DM’ mean). Thus, each participant evaluated 20 total aids (4 versions of 5 aids). Each time the participant saw a new aid, the aid was introduced as “Another hypothetical aid” suggesting that this may be a new and different aid designed to solve the same problem. In fact, this was not the case and all four modes of communicating performance were derived from the same hypothetical mean performance.
It is potentially problematic if a report with ‘DA&DM’ mean performance was immediately preceded/followed by a report of ‘DA only’ performance. It would be apparent to the participant that both reports represent an equivalent level of decision aid performance. To avoid these two stimuli appearing near each other during the experiment, all of the performance reports were organized into two blocks. For a given decision aid, the ‘DA only’ mean appears in one block and the ‘DA&DM’ means appears in the other block. All the reports within a block were presented in random order, and the blocks themselves were presented in random order. In this way, the chance of a report of ‘DA&DM’ means being presented next to its equivalent performing report of ‘DA only’ mean is minimized.

After the participant viewed the description of the decision aid along with its reported performance, they were asked to evaluate the suitability of the aid using the Likert scale for face validity used in previous experiments.

**Results- ‘DA only’ mean performance versus ‘DA&DM’ mean performance**

The first hypothesis predicts providing the mean performance of both decision maker (DM) and decision aid (DA) will result in higher face validity assessments than when only the DA mean performance is given. To test this hypothesis, two separate means were computed for each participant. The first mean is the participant’s average assessed face validity when performance is communicated in terms of ‘DA&DM’ mean performance. A second average face validity score was computed for when performance was communicated via DA mean performance only. The first mean was compared to the second mean for all participants in a paired-sample $t$ test. Two participants selected “I
cannot assess the suitability” all five times the ‘DA only’ condition was presented. Thus, no ‘DA only’ face validity score could be calculated and the data for these two participants could not be used in analysis. Data from the remaining 153 participants were used.

Opposite the a-priori hypothesis providing the additional information of DM performance hurt, rather than helped (see Figure 24), the face validity of the decision aid \((t(152) = 4.53, p<.001)\). When DA performance was provided, the average assessed DA face validity was 3.29 \((SE = .53, n=153)\). When the inferior DM performance was provided in addition to DA performance, the assessed DA face validity dropped to 3.11 \((SE = .58)\).

![Bar chart showing decision aid face validity](image)

*Figure 24.* Study 6. Decision aid face validity when decision aid performance is described via ‘DA only’ or ‘DA&DM’ conditions. Error bars represent 95% confidence intervals.
Discussion-'DA only' mean performance versus ‘DA&DM’ mean performance

Providing additional information about the marginally inferior performance of decision makers hurt, rather than helped, the perceived appropriateness of the decision aid. Participants had a higher opinion of decision aids when decision aid average performance (and only decision aid performance) was communicated. This surprising result occurred in spite of the fact that the decision aid was always better than the unaided decision makers by between 4% and 11% accuracy. Participants appeared to be unimpressed by such a small improvement in performance, and viewed the decision aid less favorably when the spread was made salient. The results suggest that, for decision aids to be viewed favorably, the improvement in performance must be greater than 11% over unaided decision makers. This represents a statistically challenging goal for an aid in a probabilistic task!

Results-Proportion of Decision Makers Beaten by a Decision Aid (High vs. low variance)

It was hypothesized that under conditions of low DM performance variance, the proportion of DMs beaten condition will be more convincing than ‘DA&DM’ condition in communicating decision aid face validity. Under high DM performance variance, ‘DA&DM’ mean performance will be more convincing than proportion of DMs beaten. To test these hypotheses, three averages were computed for each participant. The first based on the face validity assessments made by the participant when performance was communicated via proportion of DMs beaten (high variance condition). Second, the face validity when participants were given the proportion of DMs beaten when DM
performance variance is low. Third, the face validity when the average ‘DA&DM’ performance was given. These three averages were calculated for 154 participants that had data in all three modes of communicating performance (see Figure 25).

![Bar chart](image)

**Figure 25.** Study 6. Decision aid face validity when decision aid performance is described in different ways. One way is the proportion of decision makers beaten by the aid. This way has two hypothetical conditions of high and low decision maker performance variance. The other way is to give the mean accuracy of the decision aid and of the decision makers. Error bars represent 95% confidence intervals.
The three means were tested with a repeated-measures ANOVA design to test the equivalency of face validity assessments. The following a-priori contrasts were planned:

- % DMs beaten by DA (low variance) > ‘DA&DM’ mean performance
- ‘DA&DM’ mean accuracy > % DMs beaten by DA (high variance)
- % DMs beaten by DA (low var.) > % DMs beaten by DA (high var.).

The three modes of communicating DA performance did result in different levels of DA face validity ($F(2, 306) = 69.95, p < .001$). The first and third a-priori contrasts listed above were supported, while the second was not. The communication mode that resulted in the highest DA face validity was the percentage of DMs beaten by the DA in the low variance condition ($M = 3.63, 95\% \text{ CI} = 3.53 \text{ to } 3.73$), followed by the percentage of DMs beaten in the high variance variation ($M = 3.24, 95\% \text{ CI} = 3.15 \text{ to } 3.34$). Least convincing was providing the mean accuracy rate of both DA and DM ($M = 3.10, 95\% \text{ CI} = 3.01 \text{ to } 3.19$).

**Discussion—Proportion of DMs Beaten by a Decision Aid (High vs. low variance)**

It was hypothesized that when performance was communicated in terms of ‘DA&DM mean accuracy’ the mean face validity would fall between the face validity of the other two conditions (proportions of decision makers beaten: high variance and proportion of decision makers beaten: low variance). The data did not support the a-priori hypothesis. Providing participants’ with DA&DM mean accuracy was always less convincing than the proportion of DMs beaten. The ‘low variance’ condition was more convincing than the ‘high variance’ condition, but this result is somewhat uninteresting.
Participants see a high number in the low variance condition (ex: 90% of decision makers beaten) and they see a lower number for the high variance condition (i.e. 70% of decision makers beaten) so the distinction between high variance and low variance is not surprising.

Results-When the DA-DM gap is meaningful to DA face validity

The hypothesized interaction proposed that the disparity in face validity between the ‘DA&DM’ condition and the ‘DA only’ condition is much higher when the decision aid accuracy is low (see Figure 23). To test this interaction, a 2x5 ANOVA table was created with each cell containing the average assessed face validity. The rows represent the two performance conditions (‘DA only’ and ‘DA&DM’), and the columns represent the decision aid accuracies represented by the five different decision aid scenarios (arranged lowest to highest). A significant $F$ statistic for the interaction term in the repeated measures ANOVA supports the hypothesized interaction.

The interaction term (Mode x Accuracy) was significant ($F(4, 576) = 12.02, p < .001$). Plotting the means however (Figure 26) shows that the interaction was not as hypothesized (as in Figure 23). The means do not appear to diverge more as the decision aid accuracy decreases. Other uncontrolled variables, such as the size of the gap between DA and DM accuracy, and the nature of the decision scenario (do participants demand more accuracy for employee selection than election prediction?) make it difficult to interpret the nature of the interaction.
Another post-hoc hypothesis however is suggested by the results. If the decision aid scenarios are ordered not in terms of overall DA accuracy, but instead in terms of the magnitude of the spread between DA and DM performance (i.e. the difference in accuracy between the decision aid and the unaidded decision makers) a pattern is suggested (see Figure 27).

Figure 26. Study 6. Decision aid face validity for ‘DA Only’ and ‘DA&DM’ conditions. Ordered from least accurate DA to most. Error bars represent 95% confidence intervals.
When the spread between DA and DM accuracy was only 4%, face validity is lowered, rather than helped, by the presentation of DM accuracy. A spread as high as 11% does not lower the perceived validity, but neither does it improve it. One interpretation of these results is that participants thought less of a decision aid after learning that it was only, in their view, marginally better than what decision makers were able to accomplish.
To better interpret the nature of the interaction, the data were also analyzed using a regression approach. DAFV was regressed on to three main effects. One main effect included which accuracies were seen by the participant: ‘DA only’ (dummy-coded 0) or ‘DA&DM’ (dummy-coded 1). Another main effect was DA accuracy (ranging 60% to 85%). The final main effect was the difference in performance between the DA and DM (ranging from 4% to 11%). The variables were centered using the grand mean of the respective variable. The centered variables were then used to create two interaction terms that were also predictors in the regression. The first interaction term was accuracies presented (‘DA only’ or ‘DA&DM’) * DA accuracy. The second was accuracies presented * difference in performance (DA accuracy – DM accuracy). A separate regression was performed for each of 150 participants. Five participants were not included in the regression analysis because they had no dependent variable for one of the conditions (i.e. they selected ‘Unable to make a determination’ for all scenarios). The regression weights were averaged for the 150 participants. All regression weights were significantly different than zero (see Table 8).
Table 8

Predictors of Decision Aid Face Validity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean B</th>
<th>95% CI</th>
<th>t(149)=</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.29</td>
<td>[3.21 to 3.38]</td>
<td>75.72</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Accuracies Presented to Participant</td>
<td>-.18</td>
<td>[-0.26 to -0.10]</td>
<td>-4.38</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>(‘DA only’ vs. ‘DA&amp;DM’)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference in Performance (DA Accuracy – DM Accuracy)</td>
<td>.04</td>
<td>[0.02 to 0.06]</td>
<td>4.36</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>.04</td>
<td>[0.03 to 0.06]</td>
<td>10.60</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Accuracies Presented * DA Accuracy</td>
<td>-.02</td>
<td>[-0.02 to -0.01]</td>
<td>-4.44</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Accuracies Presented * Difference in Performance</td>
<td>.04</td>
<td>[0.02 to 0.07]</td>
<td>3.18</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

Note. Study 6. CI = confidence interval.

The graph in Figure 28 depicts the face validity values predicted by the equation from Table 8. It can be seen that providing the average DM performance is never helpful. A small gap between DM and DA performance harms the perceived validity of the aid. A large gap between performance appears to at best do no harm.
Figure 28. Study 6. Predicted decision aid face validity interaction plot (difference between DA and DM accuracy versus whether DM accuracy is presented or not). Where: Predicted FV = 3.29 -.18(Accuracy Presented)+.04(DA-DM difference)+.04(DA Accuracy)-.02(Accuracy Presented*DA Accuracy)+.04(Accuracy Presented*DA-DM difference). Variables are centered before inclusion in the regression equation. MGap = 7.4, MDA Accur = 72.4.

Discussion-When the DA-DM gap is meaningful to DA face validity

Surprisingly, pointing out that the decision aid outperformed decision makers appeared to hurt rather than help the perceived validity of the decision aid. The harm to face validity was especially large when the decision aid only marginally outperformed the decision maker. This post-hoc observation must be interpreted with caution however. The effect appears primarily as a result of only one decision scenario (employee
selection). The scenarios that the decision aids are embedded in may also represent an
uncontrolled confound. Perhaps greater accuracy is demanded in some scenarios (like
employee selection) versus others (like car reliability). This variable will be controlled
for in the next experiment (study 7).

Study 7: Communicating Decision Aid Performance and its Effect on Face Validity
(Second Experiment)

In Study 6, the performance characteristics (for example: 75% accuracy) were
embedded in a specific decision scenario (i.e. employee selection). The decision scenario
represents a confound. In the current experiment, this confound will be controlled by
creating 5 different versions of the questionnaire. Each version associates specific
performance characteristics with a different decision scenario, thus controlling the
scenario as a confound. Would the results from Study 6 be replicated?

H1: ‘DA only’ performance will have higher FV than ‘DA&DM’ performance.

H2: In the DA&DM condition, FV will decrease as the spread between DA/DM
performance narrows.

As a result of the previous experiment, revised hypotheses were made. The first
modified hypothesis is the prediction that adding additional information about the
marginally inferior performance of decision makers will hurt the perception of the
decision aid.

The second hypothesis explores a possible interaction effect that depends on the
gap in performance between the decision aid and decision maker. When this gap is
small, decision aid face validity will be harmed by adding information about decision maker performance. When the gap is larger (up to 11%) providing decision maker performance data will be less harmful.

Method

One-hundred and nineteen undergraduate psychology students participated for course credit. Participants were sent an internet link and were able to complete the survey on the web at a time and a computer of their choosing. Data from thirty-six participants were eliminated because, in a post-survey questionnaire, they indicated that their data should not be used because they “didn’t understand or didn’t take it very seriously.” One additional participant was eliminated for not completing the survey.

Additional steps were taken to eliminate participants who spent an unrealistically short amount of time on the survey. The time taken to complete the survey was recorded for all participants. A square-root transformation was performed to normalize the positively-skewed distribution. Seven participants were eliminated who spent less than one standard deviation below the (transformed) mean time to complete. This experiment was conducted over the internet from a participants’ home computer, and it was thought that an aggressive cut-off score of one standard deviation below the mean might be helpful since unsupervised participants might be tempted to give less than their full effort to the experiment. When this experiment took the survey and took the time to quickly read the entire question, it took 4:00 to finish. A standardized score of $Z = -1.0$ equates to a time of 3:35 to complete the 20 questions. This left 75 participants’ data for analysis.
The current experiment replicates the procedure used in Study 6 with one exception. Five versions of the stimuli survey were created that crossed decision scenario (i.e. employee selection, election prediction, etc…) with all 5 possible levels of decision aid/decision-maker performance. Participants were randomly assigned to one version of the survey. In this way, the circumstances surrounding an individual decision scenario is controlled as a potential confound.

Results- ‘DA only’ mean performance versus ‘DA&DM’ mean performance

It was hypothesized that providing additional information to participants about the marginally worse performance of decision makers would lower the perceived validity of the decision aid. For each participant (n=75), two means were calculated. The first mean was for the 5 scenarios in which the decision aid performance was provided (ex: “The decision aid was 75% accurate.”). A second mean was calculated for the five scenarios in which equivalent decision aid accuracies were presented along with information about the marginally worse performance of decision makers (ex: “The decision aid was 75% accurate. Unaided decision makers were 65% accurate”). Seventy-five percent of participants (57 out of 75 total) gave higher face validity ratings to the ‘DM only’ condition (see Figure 29).
To examine statistical significance, a paired-sample t-test compared the first mean ('DA only') to the second mean ('DA&DM'). Face validity was highest when DA performance was the only figure communicated to participants ($M = 3.25, SD = .49$). Face validity of the decision aid dropped significantly when additional information was provided about the marginally worse performance of unaided decision makers ($M = 2.92, SD = .64, t(74)=4.51, p<.001$, see Figure 30). One potential outlier participant had an unusually high difference of 2.40 in face validity between conditions. Even if this outlier is taken out, the difference between conditions remains highly significant ($t(73)=4.42, p<.001$).
Results—When the DA-DM gap is meaningful to DA face validity

It was hypothesized that communicating DM performance will be especially harmful to DAFV when DMs only slightly underperform the decision aid. The interaction is suggested in Figure 31. Notice the gap in face validity that exists between the ‘DA only’ and ‘DA&DM’ conditions. The gap is largest when the difference in performance is only 4%. The difference in FV between conditions gradually lessens as the difference in performance increases. When the difference in performance is as much
as 11%, the face validity is still harmed by communicating decision maker performance, although the harm is of smaller magnitude.

Figure 31. Study 7. Mean face validity for various differences in performance between the superior decision aid and the inferior decision makers. Error bars represent 95% confidence intervals.

To test the statistical significance of the interaction, the face validity ratings were placed in to a 2x5 table (see Table 9) where the rows represent the two methods of communicating DA performance (‘DA only’ versus ‘DA&DM’) and the five rows represent the various differences between DA and DM performance (ranging from 4% to
11%). In a within subjects ANOVA, the ‘Accuracies Viewed’ * ‘Performance Difference’ interaction term was statistically significant ($F(4, 296) = 5.82, p < .001$). In addition, the two main effects were also observed to be statistically significant (Types of Accuracies viewed: $F(1, 74) = 20.35, p < .001$; Difference in DA and DM performance: $F(4, 296) = 10.51, p < .001$).

Table 9

Mean Face Validity of Decision Aid

<table>
<thead>
<tr>
<th>Performance Gap* (DA Accuracy / DM Accuracy)</th>
<th>4%(85/81)</th>
<th>5%(60/55)</th>
<th>7%(75/68)</th>
<th>10%(70/60)</th>
<th>11%(72/61)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA Only</td>
<td>3.76</td>
<td>2.84</td>
<td>3.21</td>
<td>3.15</td>
<td>3.27</td>
<td>3.25</td>
</tr>
<tr>
<td>DA &amp; DM</td>
<td>2.99</td>
<td>2.53</td>
<td>3.03</td>
<td>2.88</td>
<td>3.19</td>
<td>2.92</td>
</tr>
<tr>
<td>Mean</td>
<td>3.37</td>
<td>2.69</td>
<td>3.12</td>
<td>3.01</td>
<td>3.23</td>
<td>3.08</td>
</tr>
</tbody>
</table>


The data were also analyzed using a regression approach. Face validity is predicted by three main effects: accuracies viewed by participants (dummy coded ‘0’ for ‘DA only’ and ‘1’ for ‘DA&DM’), difference in ‘DA&DM’ performance, and magnitude of DA accuracy. DAFV was also regressed onto two potential interaction terms: accuracies viewed*difference in DA/DM performance, and accuracies viewed*magnitude of DA performance. The continuous predictors ‘DA accuracy’, and ‘difference in DA&DM performance’ were centered using the grand means of the respective variables. Interaction terms were then created using the centered terms. This regression was
performed once for each participant using the data generated by that participant. Using this procedure, 75 estimates were obtained for each regression weight. A one sample t-test was used to evaluate which regression weights were significantly different from zero. All but one regression weight (performance gap) was significant. The analysis of the regression weights is displayed in Table 10.

### Table 10

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean $B$</th>
<th>95% CI</th>
<th>$t$(74)</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.25</td>
<td>[3.13, 3.36]</td>
<td>57.43</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Accuracies Presented to Participant ('DA only' vs. 'DA&amp;DM')</td>
<td>-.32</td>
<td>[-0.47, -0.32]</td>
<td>-4.51</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Difference in Performance (DA Accuracy – DM Accuracy)</td>
<td>0.00</td>
<td>[-0.04, 0.03]</td>
<td>-.25</td>
<td>.80</td>
</tr>
<tr>
<td>DA Accuracy</td>
<td>.04</td>
<td>[0.02, .05]</td>
<td>6.36</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Accuracies Presented * DA Accuracy</td>
<td>-.01</td>
<td>[-0.03, -0.002]</td>
<td>-2.26</td>
<td>.03</td>
</tr>
<tr>
<td>Accuracies Presented * Difference in Performance</td>
<td>.06</td>
<td>[0.02, 0.10]</td>
<td>2.86</td>
<td>.01</td>
</tr>
</tbody>
</table>

*Note.* Study 7. CI = confidence interval.

Using the regression equation to predict face validity (see Figure 32), we see a replication of the results obtained in study 6. Adding additional information about the lower performance of decision makers harmed face validity of the decision aid. Such information was especially harmful when the decision aid only marginally outperformed the decision maker.
Figure 32. Study 7. Predicted decision aid face validity as a function of DA accuracy, whether DM accuracy is presented or not, and difference between DA and DM accuracy. Where: Predicted FV = 3.25 -.32(Accuracy Presented)+.00(DA-DM difference)+.04(DA Accuracy)-.01(Accuracy Presented*DA Accuracy)+.06(Accuracy Presented*DA-DM difference). Variables are centered before inclusion in the regression equation. $M_{\text{Gap}} = 7.4$, $M_{\text{DA Accur}} = 72.4$.

**Discussion**

The results suggest providing the performance data of unaided decision makers is especially harmful when the gap between decision makers and the decision aid is small. As the gap between the superior performing decision aid and the inferior performing decision makers grows, the harm caused by providing the decision maker performance data decreases. Interestingly, at no point was providing data on the inferior performance
of decision makers helpful. Participants appear unimpressed by an improvement in performance of as much as 11%, and appear to think less highly of a decision aid if it is only slightly better (4%-5%) than unaided decision makers. The results suggest that those who are interested in promoting the use of decision aids should not advertise how much better the decision aid is versus unaided decision makers (in terms of percent improvement in accuracy of decisions) except in situations where there are only the greatest of improvements.
Chapter 6: General Discussion

Decision Aid Face Validity is a Determinant of Decision Aid Reliance

Face validity has generally been considered a topic of limited importance. This reputation comes largely from the tradition of tests and measurement design where the primary consideration is the construct validity of the measure. I have presented experimental results that suggest researchers of judgment and decision making have more reason to be concerned with the perceived validity of a decision aid. In study 1, I introduced a simple 1-item measure, adapted from Nevo (1985), to assess and quantify the perceived validity of a decision aid. The measure is short enough that decision aid researchers can easily add this as an additional data point in their studies.

The measures of face validity obtained in study 1 were shown to positively correlate with decision aid reliance for participants in studies 2 and 3. More face validity measurements were taken in study 4, and these again correlated to decision aid reliance for new participants in study 5. Participants in studies 2, 3 and 5 were not asked to evaluate the aid. They were simply presented with the aid and allowed to use the information as they saw fit. The results indicate that the participants spontaneously evaluated the decision aid for suitability, and adjusted their reliance on the aid accordingly.
Determinants of Decision Aid Face Validity

Since the manipulation was the description of the procedure the decision aid used, procedure appears to be at least one component decision makers rely on to assess validity of an aid. This is likely not the only source participants use to assess face validity. Finding additional determinants of face validity may be fertile ground for future research. Some candidates are suggested within this research paper. In study 2, the actual advice (rent estimate in dollars) provided by the decision aid was considered a confound. Subsequent studies controlled this confound by ensuring all decision aid methods ‘advised’ the same amount. Similarity of the decision maker’s estimate to the advice the decision aid provides may yet be a valid component of decision aid face validity, as suggested by the work of Davis (1996) and Al-Natoor et al. (2008).

Another component of face validity is suggested by studies 6 and 7. Communicating the performance of the decision aid appears to affect perceptions of decision aid validity. Sometimes it affects the perceived validity in ways that are not completely intuitive. Telling the decision maker that the decision aid is marginally more accurate than the subjective opinion of experts appeared to hurt, rather than help, the perception of the decision aid. Such initially unintuitive findings highlight the reason why the components of face validity warrant further study.

Applications of Decision Aid Face Validity

It was shown that face validity can define some conditional boundaries for already well established findings. In studies 4 and 5 evidence for a three-way interaction between decision maker confidence, face validity, and decision aid reliance was
provided. If decision makers are not confident, they will look to a decision aid for advice but they will also evaluate the validity of that aid. An aid that does not seem valid will not be relied upon, even when confidence is low.

Wendt et al. (2000) surveyed some of the obstacles hindering decision aid acceptance. In their paper, the authors provided a simple equation where decision aid usefulness is described as a function of relevance, validity, and work:

\[
usefulness = \frac{relevance \times validity}{work}
\]

If a decision maker perceives that a decision aid is relevant to the problem, that the approach used by the aid is valid (or face valid), and does not require too much work to implement, the aid is perceived to be useful and will be accepted and adopted. The authors discuss that it is not always easy to quantify the included terms. The current research suggests a way to quantify at least one term. Wendt’s equation, and the results of this paper suggest that improving the perception of validity of an aid can ultimately improve the adoption of the aid.

The introduction of a face validity measure provides a possible shortcut to decision aid designers interested in improving the chances their decision aid is adopted. To assess decision aid utilization, the normal path might be to introduce the aid to professionals, and allow them a period of time to use the aid if they wish. Then, utilization rates might be calculated. To improve decision aid use, the decision aid would be retooled, reintroduced, and reassessed. This represents a costly and time consuming process. A face validity measure introduces the opportunity for a designer to make minor “tweaks” to the way a decision aid is presented, and measure the reaction of such design
changes before a decision aid is fully introduced. Recording face validity is cheaper than recording actual decision aid usage rates.

Shortcomings of the Current Research

One shortfall of the current study is the brevity of the DAFV measure. The one-item measure makes no attempt to break down DAFV in to component theoretical constructs. While brevity has advantages, additional items would certainly provide additional reliability, as well as explicate underlying theoretical constructs. The current paper makes no attempt to identify what these underlying constructs might be.

Given the range of the manipulation (i.e. providing a ‘random number’ versus the advice derived from a linear regression equation) it is somewhat surprising that the observed correlations (consistently around .20) between DAFV and DA reliance (WOA) were only in the small to moderate effect size range (according to Cohen & Cohen, 2003). This is in spite of the fact that problematic WOA values were eliminated from analysis (e.g., where the final estimate did not fall between the initial estimate and the DA advice). Decision aid procedure remains only one factor that will determine willingness to adopt a decision aid. Relevance, work necessary to implement, motivation and incentives are all contributing factors.

It should also be noted decision aids in the real world will likely not have such a wide range of perceived suitability. Most real decision aids will appear at least somewhat valid, and the impact of differences in face validity scores will be smaller.
Related Research: Information Systems

Some related research from the field of Management Information Systems has examined factors that improve the adoption of technology information systems and decision support technology. For example, retail companies have developed web-based decision support technologies (called recommendation agents or RAs) that are designed to aid the customer in selecting a product. Such technologies might guide the customer to a particular model computer, digital camera, or television to purchase based on the customer’s stated needs. Companies put considerable resources toward the development of such technologies (Wang & Benbasat, 2008), so they are interested in improving the willingness of customers to rely on the aids.

The theoretical foundation of this research is that people interact with technology similar to ways in which they interact with other people. This *Computers are Social Actors* paradigm (CASA) (Reeves & Nass, 1996) supposes that even though computers and other technologies aren’t people, we judge them as if they were. Therefore, the criteria that decision makers use to evaluate a decision aid are similar to what they would use to evaluate a human expert. For example, Al-Natour et al. (2008) cite past research that shows we evaluate other individuals more positively if they are similar to us (e.g. Byrne & Griffitt, 1973). Al-Natour and his colleagues show the same is true of decision aids. We rate a decision aid more favorably if the aid uses a procedure similar to the one we would use, or if the outcome of the aid is similar to our own judgment.

Trust is an important factor in personal relationships. It is also, researchers in this area contend, important when evaluating an information system. When users interact
with a decision support technology for the first time, the user’s perception of risk and uncertainty involved with adopting the aid are particularly salient. These perceptions can only be overcome through the formation of sufficient trust in the aid (Wang & Benbasat, 2008). Trusting beliefs, in the contexts of RAs, have been broken down into three distinct components (McKnight, Choudhury, & Kaemar, 2002; Wang & Benbasat, 2008). 

*Competence beliefs* refer to the user’s perception that the RA has the requisite skills and expertise for the task. *Benevolence beliefs* refer to the perception that the RA has the user’s best interest in mind, as opposed to, say, the best interest of the company selling the product. Finally, *integrity beliefs* refer to whether the RA adheres to principles of honesty, keeping promises, and other principles a user would expect during a normal social interaction. Surveys that evaluate feelings of trust measure these three underlying constructs (e.g., Wang & Benbasat, 2008).


*Dispositional* reasons refer to a person’s general predisposition to trust others. 

*Interactive* reasons refer to assessments made of the aid’s performance and behavior during the course of normal interaction. The perceived quality of the interactive experience is influenced by the expectations the consumer has for the aid (and whether the aid meets, exceeds, or fails to meet those expectations), the control the consumer is able to exert over the aid to communicate the consumer’s needs, and the ability of the consumer to verify the performance of the aid. *Calculative* reasons refer to an assessment of costs versus benefits the aid may have to make a suggestion that runs
counter to the consumer’s interest. If the consumer perceives that there is a strong motivation for the aid to suggest a course of action that runs counter to the consumer’s desires, the consumer will have less trust in the aid. Finally, knowledge-based reasons refer to the familiarity the consumer gains with the technology. In the initial stages of interaction, information necessary for the consumer to gain familiarity may come primarily from descriptions and explanations provided by the aid itself. For example, Wang and Benbasat (2007) conducted a study showing that providing adequate explanation facilities helps to increase the consumer’s trust of the RA. Explanation facilities include information on why the RA is asking certain questions, and how it uses those questions to provide advice. Providing such explanations improves users’ evaluation of the system (Ye & Johnson, 1995; Arnold & Collier et al., 2004).

What are the differences between the research on trust in information technologies and the concept of decision aid face validity introduced in this paper? One difference is methodological. In the information technology studies, the dependent variable is a self-report survey intended to measure the participant’s perception of trust and usefulness of the decision aid. The current study instead uses observed measures of decision aid utilization. An important theoretical distinction is the Computers are Social Actors paradigm that forms the premise of the research on trust. CASA assumes that we ascribe human characteristics (like trustworthiness) to technology. Face validity is a broader concept that can incorporate characteristics not associated with human traits, such as how DA performance is communicated. In addition, the CASA research described above has focused on information technology. Some simple decision aids
could not likely be classified as a technological solution to a problem. For these reasons, decision aid face validity is a broader concept. An obvious extension of the research presented is to isolate different components contributing to DAFV. In this respect, the CASA paradigm and research on trust may fit nicely inside a larger, broader DAFV framework. Generating the parameters of this framework is important, because with low face validity, even the best decision aid is unlikely to be utilized fully.
References


Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183.


Appendix A: Study 6 Stimulus

*Predicting college success.*
Another hypothetical decision aid is used to predict whether a prospective student applying to a college will graduate if admitted. It combines factors such as test scores, high school academic performance and participation in extracurricular activities. A group of admissions officers also made predictions for each student.

**Treatment 1 (Proportion of DMs beaten by DA, low variance):**
This decision aid performed better than 98% of admissions officers who did not use the aid.

**Treatment 2 (Proportion of DMs beaten by DA, high variance):**
This decision aid performed better than 75% of admissions officers who did not use the aid.

**Treatment 3 (‘DA&DM’ Mean Performance):**
The decision aid had an overall accuracy of 70%. Unaided admissions officers had an overall accuracy of 60%.

**Treatment 4: (‘DA only’ Mean Performance):**
The decision aid had an overall accuracy of 70%.

How suitable is this decision aid for the task?
5 – The aid is extremely suitable for the problem.
4 – The aid is very suitable for the problem.
3 – The aid is adequate.
2 – The aid is inadequate.
1 – The aid is irrelevant and therefore unsuitable.
NA – I can not assess the suitability.

The treatments above are written based on the following hypothetical data, which participants did not have access to.
(M\_DA = 70\%, M\_DM = 60\%; Low variance: SD = 4.88, High variance: SD =14.93)
**Predicting Election Outcome.**
Another hypothetical decision aid predicts who will win an election 6 months before the election is set to occur. It uses a combination of poll data, trends, and the type of news stories reported in the press. A group of political pundits also made predictions for each election.

**Treatment 1 (Proportion of DMs beaten by DA, low variance):**
This decision aid performed better than 69% of political pundits who did not use the aid.

**Treatment 2 (Proportion of DMs beaten by DA, high variance):**
This decision aid performed better than 63% of political pundits who did not use the aid.

**Treatment 3 (‘DA&DM’ Mean Performance):**
The decision aid had an overall accuracy of 60%. Unaided political pundits had an overall accuracy of 55%.

**Treatment 4: (‘DA only’ Mean Performance):**
The decision aid had an overall accuracy of 60%.

How suitable is this decision aid for the task?
5 – The aid is extremely suitable for the problem.
4 – The aid is very suitable for the problem.
3 – The aid is adequate.
2 – The aid is inadequate.
1 – The aid is irrelevant and therefore unsuitable.
NA – I can not assess the suitability.

The treatments above are written based on the following hypothetical data, which participants did not have access to.
(M_{DA} = 60\%, M_{DM} = 55\%; \text{Low variance: SD} = 10.00, \text{High variance: SD} = 15.15)
Another hypothetical decision aid predicts whether the car that you are currently driving will still be running five years from now. It uses data like when the car was built, miles on the odometer, make and model, and gender and age of the driver. A group of mechanics and insurance agents also made predictions for each vehicle.

Treatment 1 (Proportion of DMs beaten by DA, low variance):
This decision aid performed better than 98% of car experts who did not use the aid.

Treatment 2 (Proportion of DMs beaten by DA, high variance):
This decision aid performed better than 78% of car experts who did not use the aid.

Treatment 3 (‘DA&DM’ Mean Performance):
The decision aid had an overall accuracy of 72%. Unaided car experts had an overall accuracy of 61%.

Treatment 4: (‘DA only’ Mean Performance):
The decision aid had an overall accuracy of 72%.

How suitable is this decision aid for the task?
5 – The aid is extremely suitable for the problem.
4 – The aid is very suitable for the problem.
3 – The aid is adequate.
2 – The aid is inadequate.
1 – The aid is irrelevant and therefore unsuitable.
NA – I can not assess the suitability.

The treatments above are written based on the following hypothetical data, which participants did not have access to.
(M_{DA} = 72\%, M_{DM} = 61\%; \text{Low variance: SD = 5.37, High variance: SD = 14.29})
**Predicting Employee Turnover.**

Another hypothetical decision aid predicts whether a prospective employee hired today will still be with the company 5 years from now. The decision aid uses factors such as employee education level, salary, and responses on a personality profile. A group of human resource specialists also made predictions for each employee.

**Treatment 1 (Proportion of DMs beaten by DA, low variance):**
This decision aid performed better than 78% of human resource experts who did not use the aid.

**Treatment 2 (Proportion of DMs beaten by DA, high variance):**
This decision aid performed better than 66% of human resource experts who did not use the aid.

**Treatment 3 (‘DA&DM’ Mean Performance):**
The decision aid had an overall accuracy of 85%. Unaided human resource experts had an overall accuracy of 81%.

**Treatment 4: (‘DA only’ Mean Performance):**
The decision aid had an overall accuracy of 85%.

How suitable is this decision aid for the task?

5 – The aid is extremely suitable for the problem.
4 – The aid is very suitable for the problem.
3 – The aid is adequate.
2 – The aid is inadequate.
1 – The aid is irrelevant and therefore unsuitable.
NA – I cannot assess the suitability.

The treatments above are written based on the following hypothetical data, which participants did not have access to.
(M_{DA} = 85%, M_{DM} = 81%; Low variance: SD = 5.19, High variance: SD = 9.76)
Predicting student success in Pilot Training.
Another hypothetical decision aid predicts whether a military student pilot will complete pilot training. The aid uses things like the student’s performance on questionnaire designed to assess flying aptitude, the student’s academic performance in college, and military performance. The student’s commander also made predictions.

Treatment 1 (Proportion of DMs beaten by DA, low variance):
This decision aid performed better than 76% of human resource experts who did not use the aid.

Treatment 2 (Proportion of DMs beaten by DA, high variance):
This decision aid performed better than 62% of human resource experts who did not use the aid.

Treatment 3 (‘DA&DM’ Mean Performance):
The decision aid had an overall accuracy of 75%. Unaided human resource experts had an overall accuracy of 68%.

Treatment 4: (‘DA only’ Mean Performance):
The decision aid had an overall accuracy of 75%.

How suitable is this decision aid for the task?
5 – The aid is extremely suitable for the problem.
4 – The aid is very suitable for the problem.
3 – The aid is adequate.
2 – The aid is inadequate.
1 – The aid is irrelevant and therefore unsuitable.
NA – I can not assess the suitability.

The treatments above are written based on the following hypothetical data, which participants did not have access to.
(M_{DA} = 75\%, M_{DM} = 68\%; Low variance: SD = 9.86, High variance: SD =22.58)