Kathy McNamara, Ph.D., Cleveland State University

The Ethics of Scientific Thinking: Assessment, Intervention & Decision-Making

Learner Objectives

Be able to identify salient ethics and legal issues in the application of RtI to assessment and intervention practices;

Learn about common myths relevant to school-based assessment and intervention;

Be able to identify cognitive errors and biases that influence school psychologists' and teams' decisions.

Ethics & Scientific Thinking

- "Ethical practice" is usually understood to mean knowledge and application of the "rules" of practice
- So, ethics training typically focuses on dilemmas encountered in practice, with solutions derived from references to such rules
- Even ethical decision-making models, which offer a systematic procedure for thinking about various aspects of dilemmas, emphasize references to the "rules" (legal and ethical codes and precedents) and how to apply them
- But "thinking ethically" requires us to first think about how we think … and to recognize that …

"The scientific method is a toolbox of skills designed to prevent scientists from fooling *themselves*"

(Lilienfeld, et. al, 2010, p. 9)

Both law and ethics require the use of evidence-based *interventions* …

- No Child Left Behind Act (ESEA, PL 89-10) regarding "scientific, research-based intervention":

  (i) Employs systematic, empirical methods that draw on observation or experiment;

  (ii) Involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;

  (iii) Relies on measurements or observational methods that provide valid data across evaluators and observers and across multiple measurements and observations; and

  (iv) Has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review (20 USC 6368).

Rank order the following factors and practices in terms of their impact on student achievement:

A. Characteristics of principals and school leaders.
B. Teachers' knowledge of subject matter.
C. Teachers' use of formative evaluation.
D. Students' socioeconomic status.
E. Students' use of self-instruction strategies.
F. Teachers' expectations for student performance.
G. Class size.

Rank order based on research (Hattie, 2009)

Both law and ethics require the use of validated *assessment methods* …

- **Individuals with Disabilities Education Act (2004 re-authorization)**
  - Evaluation must include:
    - "a variety of assessment tools and strategies"(PL 108-466§614[b][2][A]);
    - that "provide relevant information that directly assists persons in determining the educational needs of the child" (PL 108-466§614[b][3][C]);
    - and that have been validated for the purpose for which they are used (PL 108-466§614[b][3][A][iii]).

While methods may be "valid," they also must be validated for the purpose (decision) for which they are used …
So, what about IQ tests?

- Although far from perfect, IQ tests yield scores that are excellent predictors of academic achievement and job performance across just about every major occupation (even U.S. Presidents!)

- However, there is a difference between the *average* performance of African-American and White students on IQ tests, and evidence of differential performance at the *item* level ("kernel of truth"). Does this mean the tests are biased?

  - Larry P. vs. Riles (1972) decision found that a test is *unbiased* only if it yields the **same pattern of scores** when administered to different groups of people; since IQ tests yield a different pattern of scores between groups, they were judged to be biased (Bersoff, 1981)

While methods may be "valid," they also must be validated for the purpose (decision) for which they are used …
So, what about IQ tests?

- But … tests are actually biased only if they under-predict or over-predict the performance of group members on the ***criterion***:
  - Biased Test: Groups score differently on the IQ test, but perform similarly on the criterion (e.g., academic achievement)
  - Unbiased Test: Groups score differently on the IQ test, *and* perform differently on the criterion
- And even recognized bias in response patterns at the item level doesn't bias the total score.

- Therefore, the evidence indicates that the test itself is unbiased, but this doesn't solve the "outcome" aspect of decision validity considerations

- It also is important to consider …
  - Whether the use of standardized, norm-referenced tests (including IQ) contributes to disproportionate representation of African-American children in special education programs (i.e., validity of *classification of children as having a disability*, rather than of predicting their academic performance based only on currently-measured aptitude under existing educational conditions)
  - Whether group differences in IQ scores are due not to characteristics of children, but to environmental influences such as unequal educational opportunities, inadequate instruction, etc.
  - Whether placement of children in special education programs will adequately address their behavioral and mental health needs

Decision-Making in RTI/MTSS Context

- **What decisions do we make?**
  - Do tiered services (instruction, intervention) meet evidence standards?
  - Is Tier 1 instruction/behavior management achieving the agreed-upon student performance standard (what standard)?
  - Which students are at risk for failure (Tier 1 inadequate)?
  - What variable(s) should be addressed by Tier 2 and 3 interventions?
  - What intervention(s) are likely to successfully address these variables?
  - Is instruction/intervention delivered with fidelity?
  - Are students who are receiving interventions making adequate progress?
  - Given data regarding progress, what should be done about the intervention (continue, strengthen, expand/generalize, change, discontinue)?
  - Does the student have a disability? (i.e., is there evidence of an inadequate response to evidence-based interventions of increasingly greater intensity, delivered with fidelity?)

- **Are our *decisions* (not just the tests or measures we use) valid?**
  - Is a decision defensible on the basis of the technical adequacy of the methods used to make it (assessment measures, decision rules), the manner in which methods were applied, and the outcome to which the decision will lead?

Sources of Error in Making Decisions

Errors of Knowledge

Errors of Measurement and Probability Estimates

Errors of Cognition and Perception

Errors of Social Influence

Errors of Knowledge

Knowing the Evidence Base (Hattie, et. al, 2009)

Influence (Factors and Educational Practices)	Effect Size
Student self-recording/reporting of grades earned	+1.44
Effective classroom behavior management practices	+.88
Teacher clarity in delivering instruction	+.75
Feedback to students on task performance	+.75
Repeated readings interventions	+.67
Student self-verbalization and self-questioning	+.64
Direct instruction techniques	+.59
Interventions based on measured "aptitudes"	+.19
Grouping students by ability	+.12
Grade retention	-.16

Other RtI/MTSS (legal) knowledge you may not have …

(Jacob, Decker, & Hartshorne, 2011; Burns, Jacob, & Wagner, 2008)

- Parents have a right to be notified if RTI is being implemented as part of the process to determine whether a child has a disability (34 CFR§300.311[a]).

- Courts tending not to view RtI as "unreasonable delay" of special education, as long as interventions and progress are documented (Delaware College Preparatory Academy and Red Clay Consolidated School District/Delaware State Educational Agency, 2009), and "suspected disability" triggers required evaluation activities

  - May not require interventions to be implemented for a predetermined number of weeks before responding to parent request for evaluation; if request refused, must provide notice of refusal and description of rights to challenge the decision (34 CFR§300.311[a]).

Other RtI/MTSS (legal) knowledge you may not have …

(Jacob, Decker, & Hartshorne, 2011; Burns, Jacob, & Wagner, 2008)

- "Inadequate instruction" always has been a basis for ruling out a disability; RtI/MTSS screeners now provide a way of establishing the adequacy of instruction
- Consent not required (Tiers 1 and 2) to "determine appropriate instructional strategies"
  - Records review, screening, consultation
  - Interventions, if within scope of teacher's authority, and within scope of typical classroom interventions
  - Consent if ongoing involvement or privacy intrusion

True or False?

- **Adolescence is a time of psychological turmoil.**

- **Musical talent is positively correlated with IQ.**

So, many issues can be addressed and resolved in current and ongoing research, but …

- "Research can generate crucial information on … incidence, effectiveness, and consequences …
- Scientific thinking is an important *personal value* for individuals who practice psychology."
- "The evidence-based practice agenda is not just about adopting and implementing research-supported practices. It is about our way of *thinking scientifically* to reduce bias and errors in our practice" (Kratochwill, 2012, p. 38, emphasis added).

Linda is 31 years old, single, outspoken, and very bright. In college she majored in philosophy. While a student, she was deeply concerned with discrimination and social justice, and participated in equal housing demonstrations. Rank order each of the following statements about Linda in terms of their probability, with 1 = most probably true, and 3 = least probably true.

A. Linda is active in the feminist movement

B. Linda is a bank teller and is active in the feminist movement

C. Linda is a bank teller.

Errors of Measurement & Probability Estimates

Estimating Probability (Mlodinow, 2008)

- What is the probability that, in a group of 35 people, 2 of them have the same birthday?

But a probability question of greater relevance to school psychologists …

- In making judgments about progress monitoring data, how many data points are needed for a reliable decision?

Minimizing Error in Measurement: Reliability

RtI/MTSS Concerns: Reliability across CBM "equivalent" reading passages Christ & Ardoin, 2009); reliability of R-CBM "trend" (slope of trendline) with too few data points (Ardoin et. al, 2013)

- Measurement error is acknowledged as real, but we make decisions with too little regard for error
  - EG: Ratings of wine on 100-pt. scale means that wines scoring "91" are far more popular than wines scoring "87", translating to meaningful differences in sales (despite meaningless differences in quality) in this $20 billion annual business! (And this effect occurs only when numeric ratings are used)
  - A standard error of measurement of +/- 3 pts. at the 95% confidence level means that, if a test were repeated a large number of times, 95% of the time, the score would fall between +/1 3 pts. of the current obtained score … AND, in 5% of those repeated administrations, the result would fall outside that range.
- We also make many judgments based on far fewer data points than would be needed for a reliable sample … a data point might represent the mean or an outlier, and we have no way of knowing which. So it is important to know how widely the data in the distribution vary from the mean (standard deviation), as well as the standard error of measurement.

Minimizing the probability of making an error in judging student level of performance …

Table 1. Standard Error of Measurement for Grades by Reliability: CBM-R

		Estimates of reliability[a]			
		Low $r_{xx'}$		Higher $r_{xx'}$	
Grade	*SD*	.90	.94	.95	.97
First	30	9	7	7	5
Second	34	11	8	8	6
Third	39	12	10	9	7
Fourth	39	12	10	9	7
Fifth	41	13	10	9	7

Note. Standard error of measurements reported in words read correct per minute.
[a] Test–retest reliability estimates reported in the professional literature. [b] *SD* = estimates of the typical magnitude of standard deviations for CBM-R within grades

CHRIST; COOLONG-CHAFFIN (2007)

Implication: For a 4th-grade student, typical (within the average range) "wcpm" scores vary by as many as 39 points. The standard error of measurement associated with this degree of variance, if the strictest standard for reliability is employed (r=.97), is 7. So, an obtained score of, for example 120 wcpm actually means that, if the measure were repeated a large number of times, 97% of the time, the student would score between 113 and 127 wcpm … a very wide range!

Minimizing the probability of making an error in judging student improvement (slope of trendline) …

	Weeks of R-CBM Progress Monitoring (2 data pts. per week)							
S.D. of WCPM Score	5	6	7	8	9	10	11	12
8	1.68	1.28	1.02	.84	.71	.61	.53	.46
10	2.10	1.61	1.28	1.05	.88	.76	.66	.58
12	2.52	1.93	1.54	1.26	1.06	.91	.79	.69

Source: Christ & Coolong-Chaffin, 2007

Implication: Under moderately favorable testing conditions, after 5 weeks of progress monitoring (10 data pts. total), the standard error of the *slope* (i.e., weekly improvement rate in wcpm) is as high as 2.52 wcpm … quite high, especially in view of the fact that many educators regard a rate of 1-2 wcpm per week as a desirable rate of progress!

Probability matters!

- Too often, clinicians ignore research findings (expressions of probability) in making decisions about a particular case. Why?
  (1) This situation/person is unique, and an exception to the rules of probability or the findings of research;
  (2) "Probability" is irrelevant to the behavior of a specific person (which can never be precisely determined)
- But …
  - Experts routinely over-identify "counter-examples," with too great a focus on "unique" aspects and too little focus on commonalities, resulting in poor judgment accuracy (Grove et. al, 2000)
  - Clinicians' routine exposure to a sample of people experiencing more severe or persistent problems lead them to erroneously view most people as less resilient than they are, and most problems as requiring more intensive intervention than is actually needed (Cohen & Cohen, 1984)
  - And probability does matter: In a game of Russian Roulette, would you prefer to use a gun with one bullet and five empty chambers, or one with five bullets and one empty chamber?

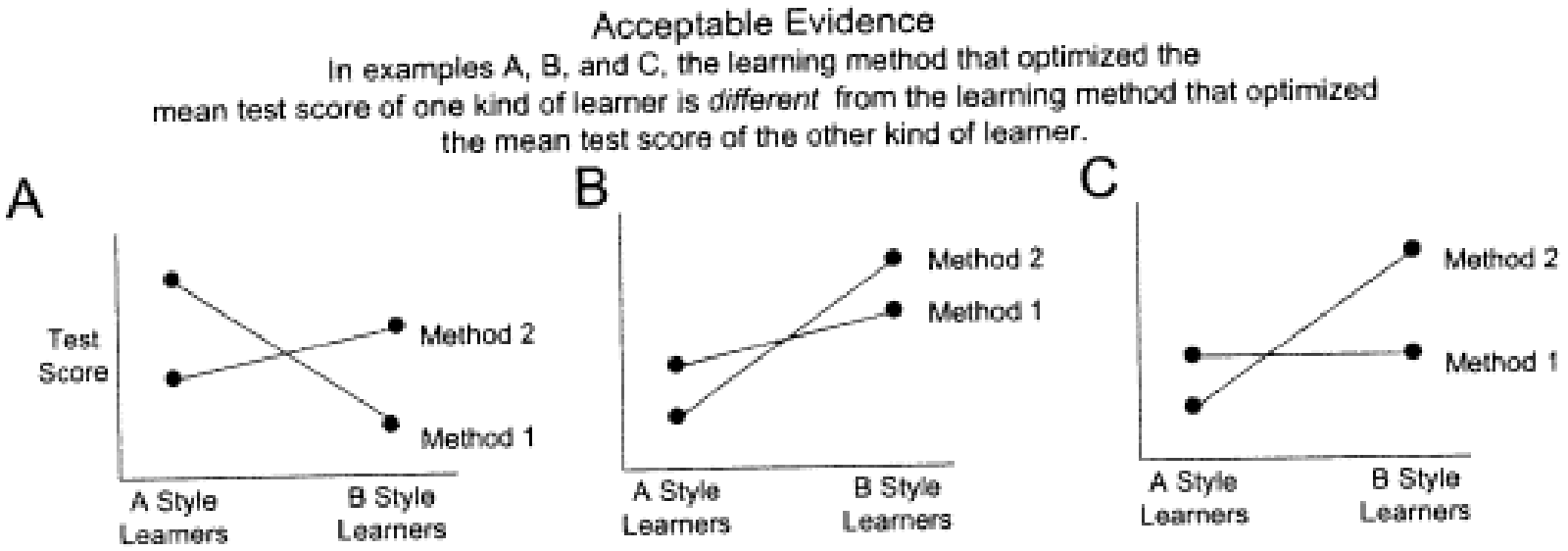
True or False?

- **Changing your answer on a multiple-choice test is beneficial only if you have a good reason to think your answer may be wrong.**

- **Matching students' learning styles to teachers' teaching styles results in improved learning.**

Learning Styles: An Illustration of the ATI Concept

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2009). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*, 105-119.

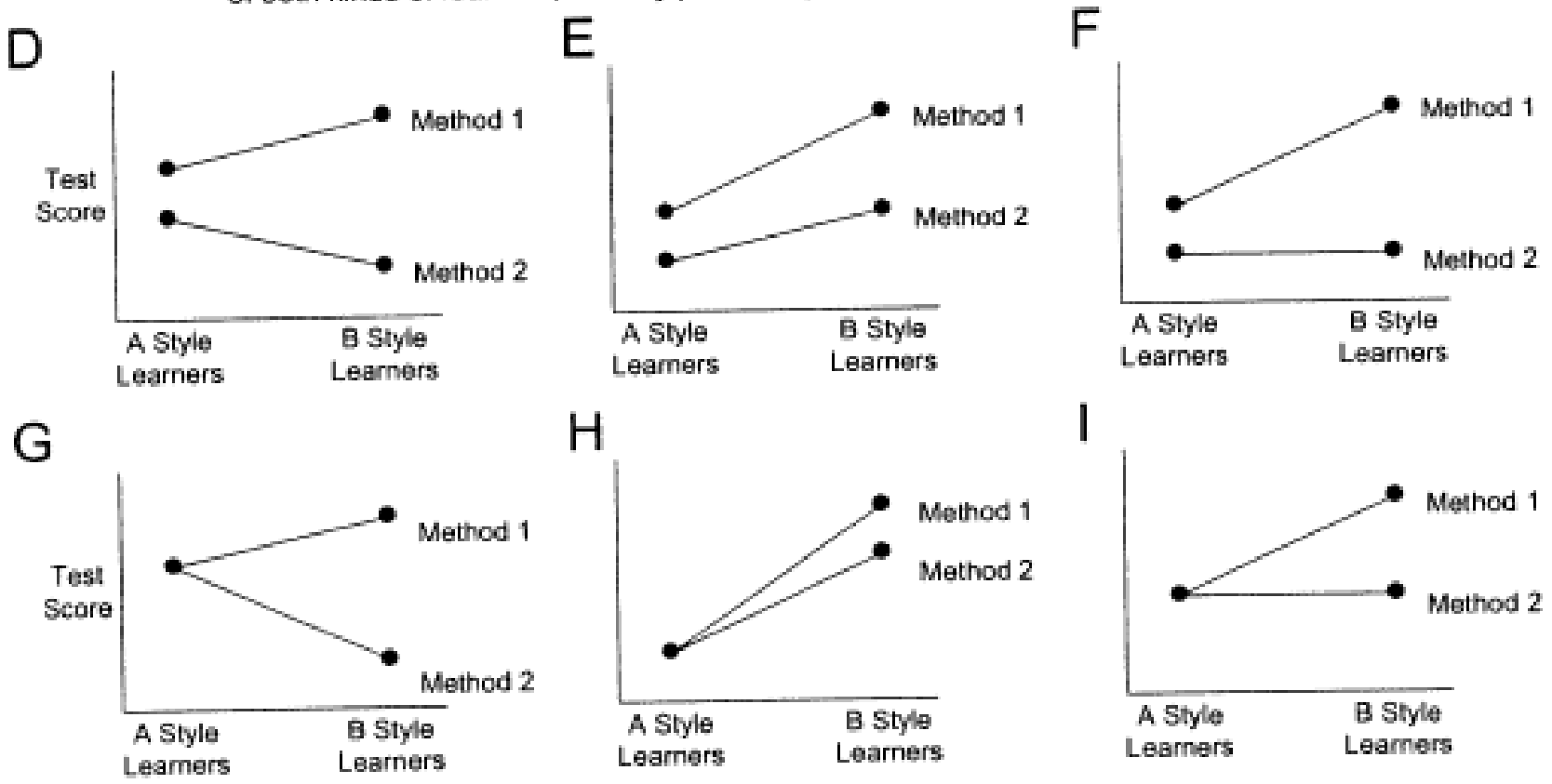


What has occurred among A & B learners, using Methods 1 & 2, in each of the above cases?

These exemplify the "crossover interaction" in which different treatment methods elicit different results in groups that differ on the aptitude of "learning style."

Unacceptable Evidence

In examples D through I, the *same* learning method optimized the mean test score of both kinds of learners, thereby precluding the need to customize instruction.



The same method maximizes the learning outcomes for *all* learners.

Inferring causation from correlation

(Lilienfeld, et. al, 2010)

- **A and B are correlated, but A doesn't necessarily cause B.**
  - The possibilities (all of which must be considered) are:
    - A causes B (maybe …)
    - B causes A (no … because the cause must precede the effect)
    - C (often unknown or unmeasured) causes both A and B (maybe …)
- Example:
  - Physical abuse in childhood (A) is correlated with aggression in adulthood (B)
  - But "cycle of violence" explanation (A causes B) although widely believed, ignores the plausible possibility of a genetic factor that "causes" both A and B (Krueger, et. al, 2001)
- Further, the "post hoc, ergo propter hoc" error frequently occurs (because A comes before B, A caused B)
  - All serial killers ate cereal during childhood, but eating cereal doesn't produce serial killers
  - And the fact that the introduction of shoes in Western culture occurred just prior to the appearance of schizophrenia doesn't mean that shoes play a role in triggering schizophrenia!

Errors of Perception & Cognition

Perceiving patterns

- Humans are “wired” to search for patterns in information, and do so reflexively
- The use of “heuristics” (mental shortcuts) conferred a survival advantage in quickly perceiving patterns, but their persistence and influence in learning, perception, and memory contributes to bias and error …
- Which, in turn, contributes to less-than-optimal decisions, particularly in situations requiring consideration of complex and sometimes conflicting information (sound familiar?)

Can you guess the underlying rule (pattern)?

- This rule is used to construct 3-number sequences.
  - The first example: 2, 4, 6
  - The second example: 4, 6, 8
- Give me examples of other sequences, and I will tell you if they obey my rule.
- Raise your hand when you think you know the rule/pattern.

Confirmation Bias

- The process of "making sense" of information often falls prey to "confirmation bias."
- Also influenced by "anchoring heuristic," in which initial impression is resistant to change

- Challenging confirmation bias …Imagine that the evidence were the opposite … would you still reach the same conclusion?
  - Example:
    - This child gets in fights with peers. She must be attention-deprived at home.
    - This child is always helping others. She must be attention-deprived at home.

- **True or False: There has been a recent, dramatic increase in the percentage of children with autism.**

- **Which is greater …**

A. **The number of six-letter English words having "n" as their 5th letter, or**
B. **The number of six-letter English words ending in "ing"**

Availability Heuristic

Judging the likelihood of an event (or accuracy of a statement) on the basis of how easily or readily it comes to mind (including believing a statement to be true if it's been repeated often enough)

Examples of beliefs influenced by the availability heuristic:

- Homeless people are more likely than non-homeless people to be mentally ill (media portrayals of or personal encounters with homeless people who were behaving oddly are more likely to come to mind)
- Gun violence, especially in schools, occurs more frequently than in the past (media reporting of school shootings leads most people to believe that gun violence is increasing, although it has decreased in the past 20 years, and is still a very rare phenomenon in schools)

Availability Heuristic

Becomes more pronounced as more time elapses between the original (recalled) experience and the current experience, if the original experience was vivid, detailed, or emotionally charged …

Which may be why educators' stories ("A school psychologist friends had a son whose hyperactivity increased to incredible levels after eating a McDonald's cheeseburger for lunch …") are more persuasive than pallid, dry research data showing no relationship between diet and activity levels or school performance!

Illusory Correlation

(Lilienfeld, et. al, 2010)

- A focus on "hits" (Table Cell A: Memorable co-occurrences) while overlooking "misses" (Table Cell B: absence of memorable co-occurrences)

  Example: Widely (and passionately) held belief that infantile autism is caused by mercury-based vaccines.

	Autism	No Autism
Mercury vaccine exposure	A ("hit")	B ("miss")
No mercury vaccine exposure	C	D

- Despite the repeated research finding of no evidence supporting association between mercury vaccine exposure and infantile autism (Grinker, 2007)

- "Post hoc, ergo propter hoc" error also may be occurring, since the appearance of autistic symptoms often coincides with the period when children are receiving vaccinations.

Hindsight Bias

- "I knew it all along" … perceiving events as more predictable after they've occurred than before they occurred
  - BUT … Although it is almost always possible to consider a past phenomenon and trace a history of events that may have contributed to the occurrence of the phenomenon, it is never possible to reverse this procedure (i.e., use current events to make predictions about the occurrence of a future event)
    - Why? Because there are so many possibilities at each step along the way, each governed by probability and circumstances, that accurate prediction is impossible.
  - Despite this, media coverage of school shootings always results in a call for screening students to find those most likely to commit acts of violence (i.e., use a limited array of current, seemingly meaningful, information to predict the likelihood of future events).

Representativeness Heuristic

- Concluding that two events (or qualities) belong together because of some superficial resemblance or quality, or because of some similarity to a past event
  - Examples:
    - Exaggerated eyes in a child's drawing is an indicator of fear, since we associate "fear" with being "wide-eyed."
    - To resolve current psychological problems, we must address their root causes in childhood … just as we must extract an infected tooth in order to cure a toothache!
    - Children from low-income homes have poor hygiene.
    - Test results indicating an average differences in performance between groups is yet another example of societal bias.
    - Verbally-oriented students will learn better when information is presented to them verbally.
    - Stomach ulcers are caused by stress (because stress causes our stomachs to churn)

Fortunately, we can count on our clinical skills and judgment (or *can* we?) …

- When supplied with the same case study information, and comparing "clinical method" (judgment and intuition applied to case data) to "mechanical method" (algorithm or "decision rule"), the latter is at least as (and sometimes more) accurate in making clinical predictions (psychiatric diagnoses, psychotherapy outcomes, suicidality, college and job performance, etc.) (Dawes, et. al, 1989)

- Malcolm Gladwell's assertions in his book "Blink" notwithstanding, studies demonstrate that intuition and "hunches" lead to poor quality of decisions in professional practice, although intuition can be a useful signal that something is amiss, and that a solution, once derived, is ethically acceptable (Cottone & Claus, 2000)

- Most clinicians think their judgment improves with experience (although it doesn't); advocate using both rule-based and clinical methods together (which works as long as both methods agree); or insist that the matter at hand is sufficiently unique as to represent an exception to the rule (which it usually isn't)… (Dawes, 1994; Grove, et. al, 2000; Smith & Dumont, 1997)

And,to make matters worse …

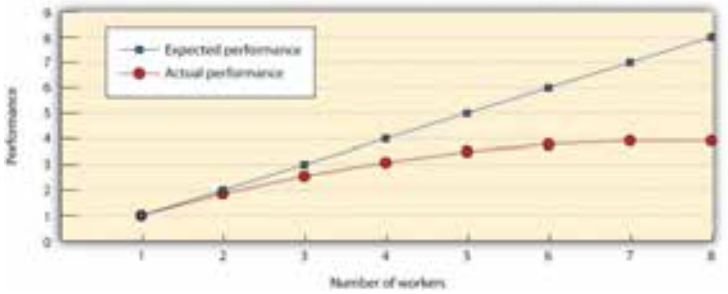
Many of the decisions in which school psychologists are involved (especially those of a high-stakes nature) are actually made by *teams*.

So, group "process" variables influence decisions, including …

Errors of Social Influence

True or False?

- "Brainstorming" new ideas in groups works better than asking people to generate ideas on their own.



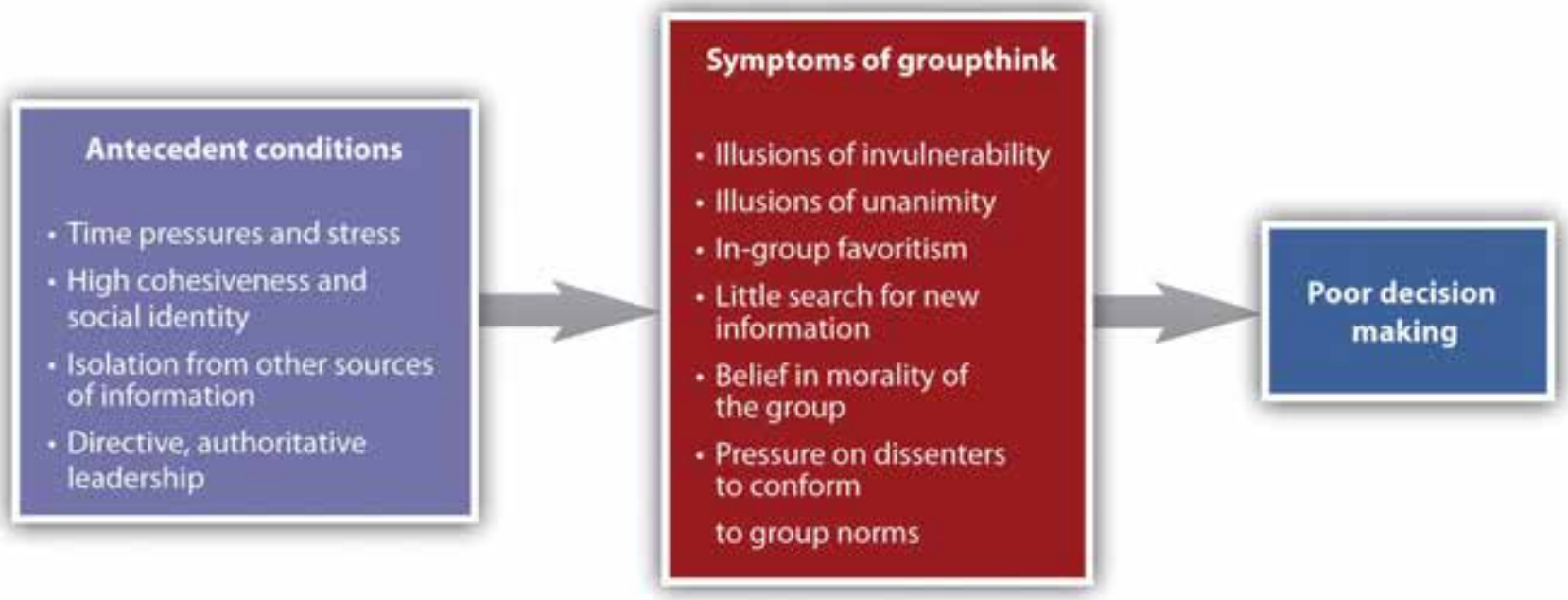
The Ringelmann Effect …

Brainstorming

- What helps?
  - "Nominal technique" (write ideas individually before the meeting)
  - "Round robin" (Sequenced turn-taking among speakers)

Groupthink

Preoccupation with group unanimity that impairs critical thinking



What helps? Actively promote minority dissent; appoint a "devil's advocate" to raise questions about group decisions.

Information sharing in groups or teams

- The three pieces of favorable information about Candidate B (b1, b2, and b3) were seen by all of the group members, but the favorable information about Candidate A (a1, a2, a3, and a4) was not given to everyone. Because the group members did not share the information about Candidate A, Candidate B was *erroneously* seen as a better choice (Stasser & Titus, 1985).

Group Member	Information Favoring Candidate A	Information Favoring Candidate B
X	A1, A2	B1, B2, B3
Y	A1, A3	B1, B2, B3
Z	A1, A4	B1, B2, B3

Furthermore, because the shared information is discussed repeatedly, it is likely to be seen as more valid and to have a greater influence on decisions as a result of its high cognitive accessibility, AND
Group members with higher status are more likely to share new information and to dominate the discussion, even if the information they have is not more important or valid (Wittenbaum, 1998; Hinsz, 1990).

Summary: Obstacles to Scientific Thinking (Lilienfeld, Lynn, Ruscio, & Beyerstein, 2010, pp. 251-252)

- Although first impressions may be helpful in "sizing up" people, they're typically inadequate when it comes to evaluating scientific claims or making decisions;
- Many shared beliefs are nothing more than "urban legends," so we shouldn't assume they're accurate;
- Media coverage, repetition, and anecdotes can lead us to over-estimate the frequency of sensational events, and under-estimate the frequency of less sensational events. Good stories aren't always accurate stories;
- Biased samples can result in equally biased conclusions. If we're exposed primarily to one group of people in our line of work, our perceptions of the prevalence of certain traits in people at large will be skewed;
- Certain biases, such as illusory correlation, confirmation bias, and the representativeness and availability heuristics, lead us to draw erroneous conclusions. Heuristics are helpful shortcuts, but if we rely on them blindly and uncritically, we'll often make mistakes;
- Correlation isn't causation, so knowing that two things are statistically associated doesn't tell us what's causing what. Also, just because one thing comes before another, the first doesn't necessarily cause the second;
- Carefully conducted scientific research (although not foolproof) is our best safeguard against error.

Ten Prescriptions for School Psychologists

- Seek out disconfirming evidence (to prove your hunch/hypothesis wrong);
- Don't become overly attached to your hypotheses ("know all theories, love some, wed none");
- Consider rival hypotheses (accept hypothesis only if it beats at least one other rival hypothesis);
- Don't cherry-pick (examine *all* evidence/data);
- Put your intuition to the test (hunches may be a good starting point, but they don't work well for decision-making);
- Be skeptical of clinical judgment and long-standing clinical wisdom ("eminence-based practice");
- Be aware of the existence of blind spots (run ideas past others to detect weaknesses or biases);
- Encourage dissent (reinforce others who offer alternative views);
- Quantify, quantify, quantify (assess "impressions" numerically; measure outcomes);
- Maintain a self-critical attitude (willingness to acknowledge that one might be mistaken), and be willing to change beliefs.

References & Further Reading

- Burns, M., Jacob, S., & Wagner, A. (2008). Ethical and legal issues associated with using response-to-intervention to assess learning disabilities. *Journal of School Psychology, 46*, 263-279.
- Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. New York: Routledge.
- Jacob, S., Decker, D. & Hartshorne, T. (2010). *Ethics and law for school psychologists.* New York: Wiley.
- Kratochwill, T. (2012). Comments on "Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error": Evidence-based interventions for grandiose bragging. *Journal of School Psychology, 50*, pp. 37-42.
- Lilienfeld, S., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50*, pp. 7 – 36.
- Lilienfeld, S., Lynn, S.J., Ruscio, J., & Beyerstein, B. (2010). 50 great myths of popular psychology. West Sussex, UK: Wiley-Blackwell.
- Mlodinow, L. (2008). The drunkard's walk: How randomness rules our lives. New York: Vintage.
- Nisbett, R. & Ross, L. (1980). Human inference: Strategies and shortcomings. Englewood Cliffs, NJ: Prentice.
- Watkins, M. (2009). Errors in diagnostic decision-making and clinical judgment. In T. Gutkin & C. Reynolds (Eds.), The handbook of school psychology (pp. 210-229). New York: Wiley.