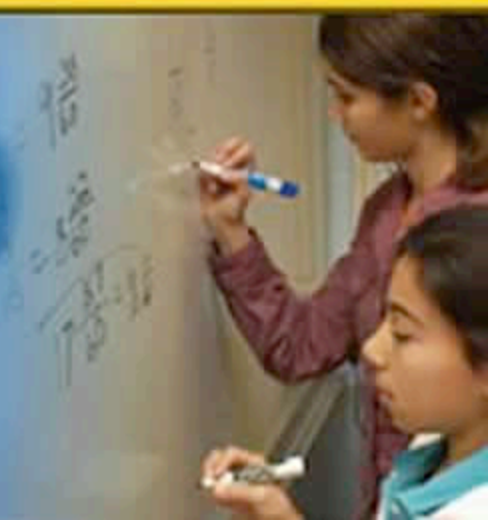
National Mathematics Advisory Panel

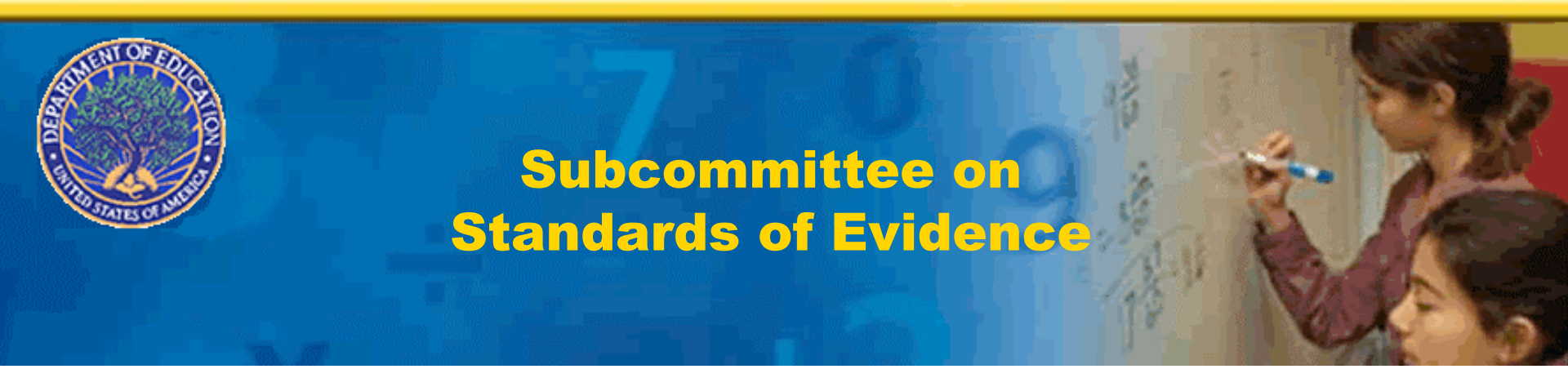
Subcommittee on Standards of Evidence

St. Louis Meeting

Progress Report

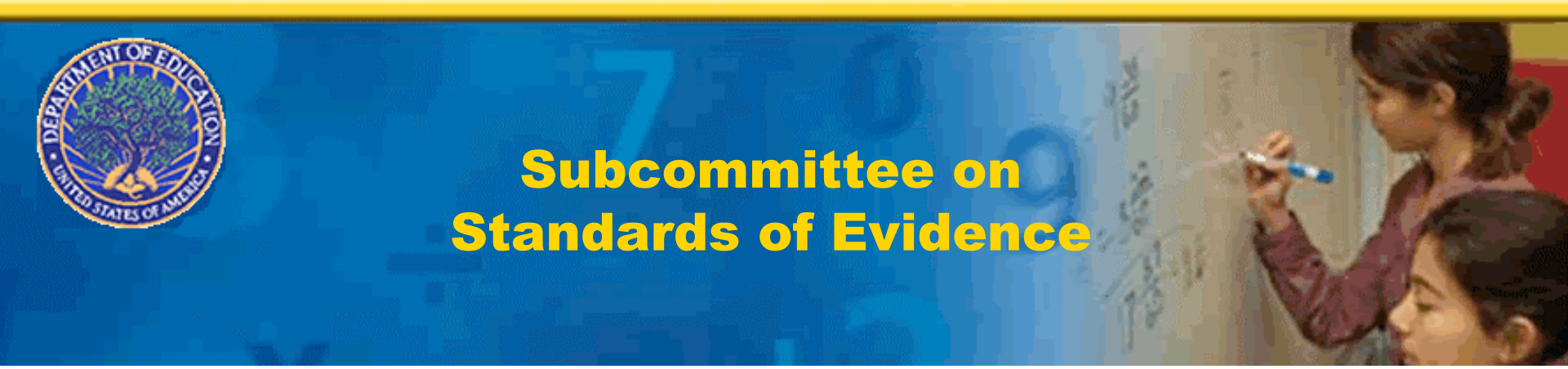
September 7, 2007





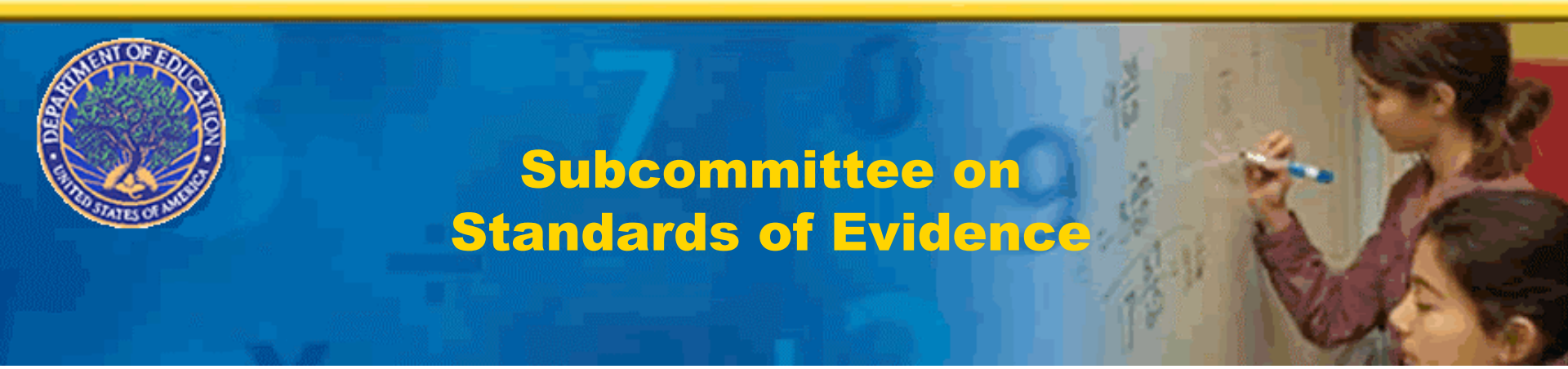
Subcommittee on Standards of Evidence

- Executive Order
  - Marshal the best scientific evidence
  - Evidence-based mathematics instruction
- What is the best scientific evidence?
  - 3 broad categories of quality
    - Highest quality = high internal and external validity
    - Promising or suggestive = has limitations
    - Opinion = values, impressions, or weak evidence



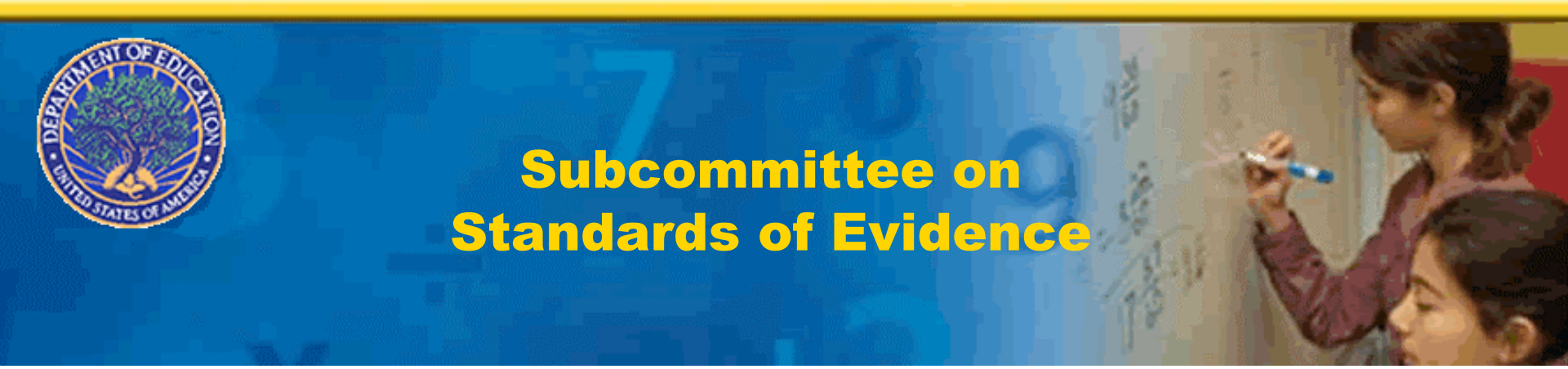
Subcommittee on Standards of Evidence

- Strongest confidence for studies that:
  - test hypotheses,
  - meet the highest methodological standards (internal validity),
  - have been replicated with diverse samples of students under conditions that warrant generalization (external validity).



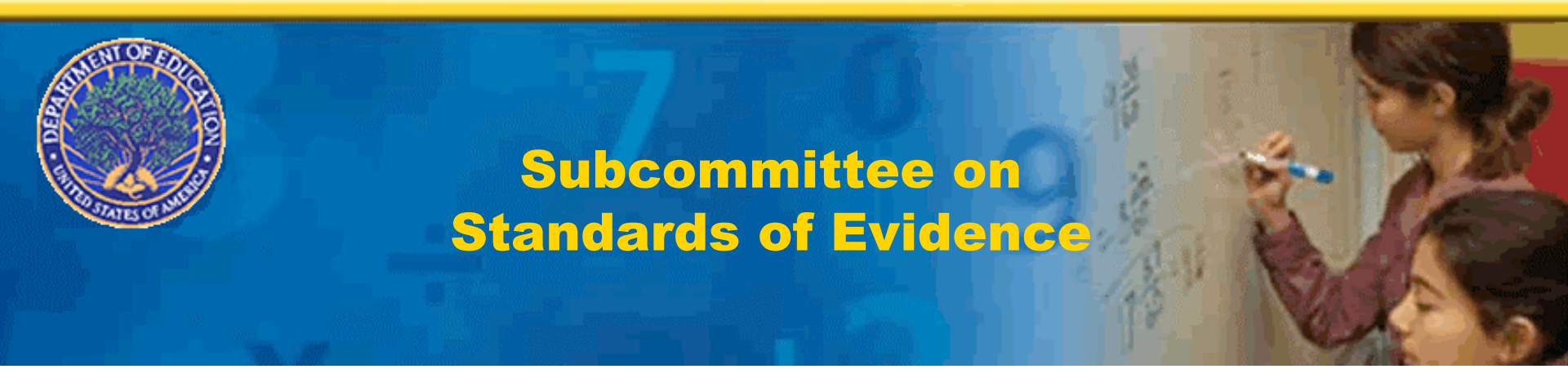
Subcommittee on Standards of Evidence

- ***Strong Evidence***
  - All high-quality studies support a conclusion
    - Statistically significant individual effects, significant positive mean effect size, or equivalent *consistent* positive findings
    - At least 3 independent studies with different samples and settings or 1 large high quality multi-site study
    - Any studies of less than high quality show either a preponderance of consistent evidence (e.g., mean positive effect size) or such weaknesses that they do not provide credible contrary evidence.
  - Factors such as error variance and measurement sensitivity influence the number of studies needed to support a conclusion (Killeen, 2005)
    - Number and balance of studies indicated are rules of thumb (e.g., see evidence standards applied by the What Works Clearinghouse at www.whatworks.ed.gov).



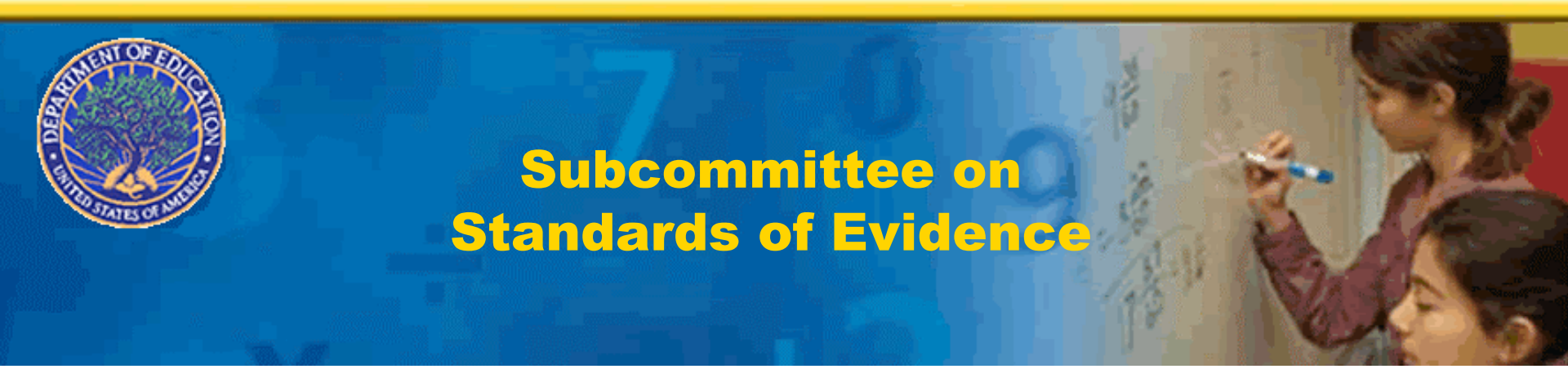
Subcommittee on Standards of Evidence

- ***Moderately Strong Evidence***
  - 1 or 2 high quality studies, or effects not independently replicated by different researchers, or do not involve different samples and settings.
- ***Suggestive Evidence*** (1 of the following):
  - Some high quality studies support the conclusion (statistically significant effects, significant mean effects) but others do not (null, not significant negative effect).
  - No high quality studies, but all moderate quality studies support the conclusion (statistically significant individual effects, significant positive mean effect size, or equivalent consistent positive findings) and there are at least 3 such studies.
- ***Inconsistent Evidence***
  - Evaluation of mixed evidence depends on quality of designs and methods. *Results of high-quality designs trump inconsistent or null results of low-quality designs.* Mixed results of high/moderate quality studies not in above categories.
- ***Weak Evidence***
  - Only low quality studies are available.



Subcommittee on Standards of Evidence

- Standards for quality differ for different kinds of research and inferences (Shavelson & Towne, 2002).
  - Effects of interventions:
    - *High quality*. Random assignment to conditions; low attrition; valid and reliable measures.
  - Descriptive surveys of population characteristics:
    - *High quality.* Probability sampling of a defined population; low nonresponse rate or evidence that nonresponse is not biasing; large sample (achieved sample size gives adequate error of estimate for the study purposes); valid and reliable measures.
  - Tests and assessments:
    - Psychometric standards such as measures of validity, reliability, and sensitivity (e.g., Anastasi, 1968; Cronbach & Meehl, 1955).



Subcommittee on Standards of Evidence

- Systematic reviews yielded hundreds of studies on important topics, but only a small proportion met standards for high quality.
- Many failed to meet standards because they do not permit strong inferences about causation or causal mechanisms (Mosteller & Boruch, 2002; Platt, 1955).
  - Rely on self-report, introspection about what has been learned or about learning processes, and open-ended interviewing techniques despite known limitations (e.g., Brainerd, 1973; Nisbett & Ross, 1980; Woodworth, 1948).
- Therefore, the Subcommittee recommends that the rigor and amount of coursework in statistics and experimental design be increased in graduate training in education.
  - Such knowledge is *essential to produce and to evaluate scientific research in crucial areas of national need*, including mathematics education.