Statistical Critique Summary
Prepared for NOMMA

by Whorton Marketing & Research

Contents

Abstract

We have reviewed a number of articles citing research results concerning the "climbable guards" issue. Our position is that many design modification proposals for guards are well intentioned, but they have not been based on statistically reliable, evidence-based arguments.

Applying conventional standards used in social science research, past articles fail by a considerable margin, and failed by a greater margin if we applied more rigid standards often required by testing in physical sciences.

In part because NOMMA members and the industry are less oriented toward arguments based on emotion or unreliable manipulations of primary and secondary research, and are more familiar with the rigid standards which routinely applied to the manufacture and installation of products, they have maintained for some time that definitive, reliable research should be conducted that will support or reject the development of new standards and codes. Over the past decade they have responded in a generally reactive way to the arguments that have been published and presented by various proponents.

The purpose of this document is to discuss primarily from a statistical perspective the weakness of past research cited, and to support other writings advocating a specific experimental design that would allow the industry to develop a reliable information base that would allow them to measure the possible public health and safety effects of new codes and standards.

A. Criteria Used for Evaluating Studies

Our purpose in this critique is to analyze studies that have been cited in the past prominently to support code changes. As with other aspects of our review of the past research conducted in this field, we encounter difficulty establishing standards and engaging in a robust critique of the research. Even the most factual presentations and articles which seem to be grounded in research actually use available studies to support a position.

Hypothesis Testing

In the research that has been conducted there are generally no clearly stated hypotheses for testing. A hypothesis is typically a clear statement that precedes the study's design and execution phases, which can be proved or disproved by the data collected. A relevant hypothesis for the matter at hand might be worded "children are able to climb the specific Design X at a significantly higher rate than the alternative Design Y" or the far higher standard "Design Y is proven to be unclimbable."

Note that most hypotheses need not be specifically concerned with a child's motive or interest in climbing as a direct factor driving the likelihood of a fall/jump accident incidence. However, the standards of hypothesis testing do require sufficient numbers of observations to ensure that the findings are statistically significant and not an odd, random occurrence that cannot be replicated and does not hold true for the population at large. This in turn requires more iterations of experiments than we are accustomed to seeing in the literature, a number that might be impossible to achieve if additional elements were introduced into the study design such as simulation of greater hazards such as a visual cliff that would reduce the likelihood that subjects would actually undertake the action being tested.

Replicability

One major standard of scientific research is that it should be possible to replicate the experiment and yield the same outcome under these circumstances. Because replication of studies can be expensive, it is not done often and acceptable substitutes include clearly communicating all protocols, features and criterion used in the study (at least leaving the researcher open to having their study confirmed or denied), and by the common practice of using secondary databases developed by reputable third parties such as government agencies or conducting literature reviews of a cross section of past studies.

This standard might be expressed more conversationally as "in a world where we could either magically repeat the experiment 100 times or observe the behavior of everyone in our universe, we could reliably predict that Design Y does outperform Design X in terms of safety, or that no individuals could climb Design Y."

Relevance/Applicability of Findings

Beyond the issues of statistical reliability, the findings of research must also be applicable to the real world. Here as with most subjects, the applicability of research to commerce and markets becomes far more murky. In other writings we have drawn comparisons to other areas of product liability where there are published studies that tend toward the status quo due to the weakness of secondary data sources and the tendency for primary research studies that violate construct validity by leading to ambiguous interpretations. There is also the basic issue of fields where regulatory decisions cannot be pursued because the inherent nature of the event we seek to prevent—in this case fatal or serious injuries from falls—occur infrequently and cannot be simulated, leaving us without valid secondary data sources, and only adequate prospects for conducting reliable, definitive primary research.

To use an analogy, we know that guns can be dangerous if accessed by children. We might even conduct a test to prove that submachine guns are more likely to cause fatalities among children than are, say hunting rifles. However, there has been little testing of this and generally research has not affected any standards governing the specific regulation of gun designs. In part this is because all guns are regulated heavily, child safety is just one of a variety of public health and safety concerns considered in the regulation of these products, and the incidence of accidents involving the individual products occurs too infrequently to warrant extensive specific action.

This is not to say that there isn't considerable debate raging on the issue, from individual and large organizations who advocate specific stances. However there is no research conducted and regulatory bodies do little in these areas but monitor the debate for new arguments and working within the evidence based standards which limit the flexibility they have to pursue solutions for which actual outcomes are difficult to simulate under experimental conditions.

Expansion/Growth of Knowledge

In most fields of applied research, studies are designed to build on prior studies, such as examining a key point/principle which was left untested in previous published research, or determining if the study can be applied to alternative experimental conditions. The primary research studies we have reviewed generally repeat the same issues, sometimes emulating other studies with near-identical conditions. They are also not designed to prove or disprove specific assertions that have been made with policy implications (such as those advocating outright elimination of certain designs through the code process). Thus far the research that has been conducted seems to have more adversarial objectives in that they have been conducted without independent review of the study design by all parties prior to conduct. We seek to propose a far more open approach to study design by building on past studies, seeking to disprove or prove them in a reliable way that meets the requirements for scientific, evidence based arguments despite any subjective individual perspectives on the issue.

Resource limitations are another reason for taking this more open, transparent approach to study design. Study design and execution can be expensive, and the cost of poor design would be lost resources and also lost time as additional studies would be required and any appropriate code or regulatory changes would be delayed by design flaws, such as inadequate sampling and absence of conditions such as a visual cliff, similar to the weaknesses we have seen in past studies.

Representative Sampling

Most of the evidence that has been cited, with considerable effort by all parties involved, has been anecdotal. Press accounts from a mix of international and domestic sources always represent a single data point, one that has reflects an accidental occurrence. Even in our initial engagement for NOMMA we intended to compile a full database of pertinent accident reports that have been reported through the world's press but abandoned this effort for being too many steps removed from our actual purpose, which is to begin planning and to advocate for a structured program of research that meets the common needs of all parties involved—to definitely and accurately measure the impact of codes and design on the public safety, generally to be measured by the probable incidence of child accidents in the form of falls. The press accounts taken as a whole, no matter how exhaustively cataloged, nonetheless serve as a convenience sample. They provide anecdotes for stories to illustrate various points, but they do not make the point or prove anything.

Literature Surveys
Often when there has been considerable research conducted, a literature survey is conducted to review and summarize the studies that meets key standards. Fields vary according to their intellectual rigor, but it should be noted that in areas such as medical literature the standards are very high. A sample literature review will critique studies and only cite findings if the study uses a set acceptable analytical techniques (such as analysis of variance [ANOVA] or multiple regression; appropriately cites their sample selection characteristics (including stated criteria, any drop out or loss of data due to inability to follow up with subjects). The studies will be rated as good, acceptable or unacceptable and may be scored for overall methodological quality.

We cite this precedent primarily to note that public health studies typically should conform to the higher standards of medical or epidemiological studies. There is no substantial body of literature available, nor are there well defined criteria as there are established by investigational review boards (IRBs) for study subjects. This reflects the rarity of the industry dynamic we are facing, and it also explains why there has been serious consideration given to what is very weak evidence. We respect the effort that has gone into identifying and presenting data, and it's accepted as the best available information in the absence of anything better that would provide sufficient, statistically reliable predictive data.

Controlled Environments
Any experimental study conducted in the past should have a fully described methodology and controls applied to ensure reliability. In industries that rely on human participation, such as food and drug testing, there are clear rules for subjects of research. For example, FDA regulations must be satisfied for approval research projects that will support future claims of product safety or efficacy. Criteria for initial review and continuing review include factors such as risks to subjects being minimized and reasonable in relation to anticipated benefits; selection of subjects is equitable; informed consent is adequate and appropriately documented; research plans makes adequate provisions for monitoring data collection to ensure safety and privacy/confidentiality of subjects; and appropriate safeguards have been included to protect vulnerable subjects. [21 CFR 56.111]

B. Specific Critique

To better illustrate our criticisms of past analysis and to support our recommended standards for the research that will be conducted in the near future, we choose one recent synopsis—Elliott O. Stephenson, *"Climbable Guards: The Special Enemy of the World's 2- and 3-Year-Old Children"* The Code Official, January/February 2002, BOCA International Inc. Pages 36-41. This article cites a number of studies and we discuss each of them below with a summary and some observations from the statistical/scientific perspective.

1) Netherlands Study—In a study conducted by the Netherlands' Consumer Safety Institute, a total of 66 children helped to test eight distinct combinations of assemblies and guard heights. Descriptive data was provided regarding the ability of children to climb each configuration. Most of the summary cited here concerned 17 children who were in the age ranges from 2½ to 3½ years. The summary indicates that between 63% and 80% were stopped from climbing by specific guards.

- A rigid guard design with rigid verticals spaced 6" apart and a 39.3" high assembly, stopped all test subjects up to 5 years but failed to stop between 20% and 40% of older children.
- A flexible guard with flexible top wire 47¼" and an 8" platform in front of the guard could not be climbed by children under 3½ years but failed to stop between 20% and 40% of older children.
- Other designs including a wire fence guard with solid top rail, a guard with a taut top wire between rigid posts, and a welded panel guard, each stopped a limited number of children in most of the age ranges examined.

Observations/Conclusions—This study seems to document that some guards are climbable—an assertion we previously accepted. The study, which was published in the *International Journal for Consumer Safety*.

- This study highlights the weaknesses of applying systematic product testing data without a specific hypothesis. We are unfamiliar with specific features of Dutch liability laws and current codes, hence the reason for testing, but there are no conclusions drawn here regarding the superiority of one or more designs, or the unacceptability of any of them. Because no guard could prevent climbing with all the age cohorts summarized in the study, we might conclude outside the scope of this study that no guard is successful in preventing a child from willfully climbing and thus jumping or falling in an accident.
- If the goal of the study was to establish linkages between behavior and guard design, there were no provisions to ensure statistical reliability The sample size is inadequate—the limited number of participants overall and particularly by age cohort make it impossible to calculate a confidence interval around the estimated percentages who can climb specific designs. That is, not only can we not determine if the proportion of children in an age range would be stopped 20% or 40% of the time, we cannot project an estimate less than plus or minus 20% even at the traditional 95% level of significance used in most studies, or a far less restrictive 90% level!
- This study highlights findings which summarize the number of children at various ages who can climb a variety of guards--presumably with some incentive or encouragement present in the study to attempt to climb. As with some other studies we see photographic images of children, sometimes smiling. While it sounds callous to point out, because all barriers are at ground level, we again are missing the vertical distance that can make some falls dangerous but also reduce the incentive for a child to climb. The formal definition of "climability" (a term that some have objected to in literature and the overall CTC charge) for any installed guard/barrier should include facilitating the physical ability to climb and the psychological propensity to do so.

2) New Zealand Study—A study commissioned by the New Zealand Building Industry Authority tested nine guard assemblies with a total of 24 children. This study also attempted to measure the likelihood that children could climb the guard designs and described the physical actions taken to climb the designs.

- All designs were 39.3" high and each included a solid top rail and included full or partial solid panels and either spaced vertical or horizontal bars.
- The ages ranged from under one year to just under five years. All designs were climbable by at least one of the participants, or 4% of the sample.

Observations/Conclusions—This study documents that all guards are climbable and the sample size is even smaller, yielding even broader confidence intervals that make it impossible to replicate the data. The variance in results (proportion of human subjects able to climb the guards) across these studies also tend to reinforce the practical matter that the study results cannot be replicated in subsequent studies, because the initial findings are unreliable.

3) Brisbane Australia Study—This study has 515 children between 2 and 8 years tested on barriers between 24 and 54 inches.

- The barriers tested were pool fences with designs that offered a vertical toe hold every 36 inches.
- Results indicate that 50% of 3 year olds could climb a 36" barrier and 20% could climb a 48" barrier.

Observations/Conclusions—This study has a sufficient sample size to ensure that at some reasonable estimates can be made for the climbability of specific designs. Assuming roughly equal distribution of children by age, the confidence intervals for the proportion of a cohort such as 3 through 4 year olds might be as narrow as +/- 9%. Therefore, the study finding that 20% can climb a 48" barrier with the specified design means that in reality, between 11% and 29% can climb it.

- This study's real limitations stem from the fact that results are "directional" but not directly applicable to the U.S. environment. These designs tested reflect pool fences only of varying heights but with otherwise consistent designs.
- Applying appropriate controls to the study would entail having a series of designs that are of concern to the CTC and/or are covered under current code proposals, to have each child in the sample climb each one in turn, and to test for statistically meaningful differences in the climbability of a specific "acceptable" design and one or more "unacceptable" designs.
- Although it is unclear from documentation whether this study included such necessary design features, a U.S. based study building on these findings would have to include additional provisions to simulate a natural environment. This would allow psychological influences such as the presence of a visual cliff that would allow study investigators to also factor in the desirability and thus likelihood of actual climbing by child subjects within specific age cohorts. Since various designs may also affect a child's willingness to climb, representing a behavioral choice that would balance the appeal of climbing vs. some alternative use of time, inclusion of external stimuli is an essential feature to test the hypothesis that specific designs facilitate or deter climbing related accidents.

4) NEISS Data—The National Electronic Injury Surveillance System data is reviewed to provide some perspective regarding U.S. accident rates. The CPSC released data from NEISS covering the period of January 1994 through mid October 1999 which we have critiqued in detail with other reports to NOMMA.

Although the data is reported from only 101 hospitals, our review of their profile indicates that the article's practice of projecting the sample nationally by a randomly chosen factor of 40 to represent all accidents reported in the 5,400 hospitals.

The government's own statistical reports have to use a complex weighting scheme to account for their oversampling of small and children's hospitals, and we do not know how representative this sample is of all U.S. hospitals.

- The universe of community hospitals in the U.S. is heavily weighted toward smaller institutions, with almost 50% of hospitals having below 100 registered beds, but they make up only 12% of total U.S. hospital beds and thus, probably a small proportion of total hospital visits.
- NEISS intentionally oversamples smaller hospitals (47 of the 100 hospitals are below 16,000 annual emergency room visits) and children's hospitals (5 of the 50 total in the U.S. are in the sample).
- Any statistical weights used for projection across the U.S. must be very substantial.
- Incidents reported in children's hospitals will not be representative of national activity: they account for 1% of all national hospital incidents and roughly 2% of all emergency room visits.
- We would also need to ensure that Incidents reported in smaller hospitals are unbiased and representative of events occurring in other larger hospitals. Because of the oversampling of smaller hospitals, their incidents would require a weighting factor that is roughly 1/10th of that reported in the larger hospitals of the NEISS sample to be projected nationally.

Observations/Conclusions—Because of these reasons, we believe it is best to consider the data that is observed and not attempt to project it to represent the entire U.S. without a far broader data collection effort that would ensure reliable national estimates. To do so with the existing NEISS sample to estimate the incidence of accidents and to support hypotheses regarding them is to twist the data in a manner unsupported by the government agency responsible for collecting and reporting the data.

	Study Period		Annualized		
NEISS Hospital Reports	**Falls**	**Jumps**	**Falls**	**Jumps**	**Proportion**
Balconies/upper levels	313	34	55.2	6.0	14.7%
Porches and decks	1159	217	204.4	38.3	58.3%
Banister or rail (from or off)	414	29	73.0	5.1	18.8%
Banister or rail (over/onto)	176	18	31.0	3.2	8.2%
Total	**2062**	**298**	**363.6**	**52.6**	

- Eliminating the projections and presenting only the Actual Reports data from Page 37 of the article tells a different story. The data is collected and reported for a multiyear period to control for seasonal factors and anomalous periods, but the annualized incident statistics show that in the NEISS sample of participating hospitals, an average of 363.6 falls and 52.6 jumps generating hospital visits were reported per year. This translates to one fall accident per day, which tells a very different story and suggests that accidents are rare. Coupled with other data regarding the low proportion of severe accidents, even an acceptable method of projecting to a national sample corroborating with existing published sources of information reinforce the impression of rare incidents that may not be reduced at all through implementing and enforcing code changes.
- Using accident reports (or later as we will examine, hospital admissions) tend to sharply overstate the problem because the category of accidents includes a heterogeneous wide range of incidents that led to an injury outcome which in turn varies in severity considerably

from incident to incident. This variance can be difficult to quantify in a way that is not subject to further debate, because unlike other medical phenomena such as disease entities and procedures for which the ICD9 international classification system exists for reporting statistical information in medical billing, there is no similar taxonomy or classification scheme for the severity of injuries that is universally applied. So we must look for proxies for all accidents and extrapolate their distribution by probable severity.

- For corroborative evidence, we cite the comprehensive *2004 State of Home Safety in America*. This report was prepared with very high standards to cover a wide variety of residential accidents. The report cites that the ratio of nonfatal home injuries to every home injury death is 650 to 1. For falls the ratio is 860 to 1. The report indicates a total of three deaths per year. Because of the small number of fatalities that occur annually, we cannot draw long-term conclusions regarding how susceptible children are relative to other individuals—each accident is a unique occurrence and an increase from 1 to 3 in a year could be cited disingenuously as "a 300% increase" just as a decrease from 3 to 1 in the next year could be cited as a "67% decrease." As with all summary statements regarding statistical trends, we must be cognizant of the denominator uses in these analyses.
- In one key area, advocates for child safety may be under representing overall considerations of home safety by focusing purely on children. According to the *2004 State of Home Safety in America*, children 1-4 are actually less likely than elderly 80+ to experience a fall in the home. The younger age cohort actually comprises only 10% of all home injuries. A comprehensive analysis of accident statistics and consideration of the impact of product design would have to take into account the effect of barriers on a variety of age ranges. We would not advocate an experimental design that tests the ability of older individuals to climb guards, and this is perhaps the most damaging critique of the past experimental research—just because a person lacks judgment and climb something does not create a product liability or cause for greater regulatory restrictions on the part of manufacturers and installers.