METHODOLOGY

BACKGROUND AND OVERVIEW

This study identifies and analyzes the messages involving sex and sexuality that are presented across the overall television landscape. Three complementary sampling strategies were employed to obtain the programming examined for the study. First, a composite week for each of ten of the most frequently viewed channels representing all aspects of the television industry was gathered by randomly sampling programs between October 2004 and April 2005. Second, an over-sample of broadcast network prime-time programming was also collected during the same period. And third, a sample of the programs most frequently viewed by teenagers, an audience of particular interest in this realm of potential media influence, was gathered. Programs for each of these three samples were digitally recorded and then systematically evaluated using scientific content analysis procedures applied by trained coders.

A total of 1,154 programs were analyzed for the project as a whole. In this description of methods, we provide complete details about the process for sampling programs, the nature of the measures used to describe and evaluate the content, and the consistency of coders' judgments in analyzing the programming.

SAMPLE OF PROGRAMS

One of the key goals of the study is to produce findings that can be generalized to the overall television environment. We rely upon the composite week sample, which is our primary focus of attention in the report, to achieve that goal. The channels included in the study were selected to encompass the full diversity of competitors within the industry, including commercial broadcast, public broadcast, basic cable, and premium cable channels. The composite week sample is highly representative of the full range of content that appears on television, with some modest exclusions which are detailed below.

Because of our particular interest in prime-time broadcast network shows, a separate "over-sample" representing three full weeks worth of the nationally distributed evening programming on each of the four major networks was also collected, as indicated below. Increasing the depth of the prime-time sample allows us to have greater confidence in the findings we report regarding these prime-time network shows, which despite their recent decline in audience share still remain the most heavily viewed programs on television.

Finally, our third sample consists of three episodes of each of the 20 most frequently viewed television series for those between 12-17 years of age, as determined by the national audience ratings for this age group by the A.C. Nielsen Company. In reporting our data, we always specify which sample of programming is involved: the overall composite week, the network prime-time over-sample, or the teen program sample. We never combine these groups of programs for any analysis.

The method by which programs were selected for each of these three sample groups, as well as the implications of these procedures for ensuring strong generalizability of the findings, are presented below. We first review the composite week design.

Composite Week Sample

For each channel included in the study, a composite week sample spanning the 16 hours daily between 6:00 a.m. and 10:00 p.m. Mountain Standard Time (MST) is constructed by a procedure of random selection. This process begins with an empty grid of half-hour time slots for all seven days of the week for each one of the channels studied. Then, across a span of approximately six months (October 11, 2004 to April 18, 2005) that comprised the sampling period, half-hour time slots are randomly selected for recording. This process yields a collective total of 112 hours of programming per channel.

Once a time slot and channel are identified, the upcoming week's TV Guide is checked and the corresponding program is scheduled for taping and placed on the sample grid maintained for each channel. Programs extending beyond their half-hour time slot are recorded and analyzed in their entirety, and placed on the grid accordingly. Appendix A presents the complete list of programs sampled for the composite week.

With the random selection process, each program that airs has an equal chance, or probability, for inclusion in the sample. Because random selection assures us that each program is chosen independently from all the others, we can be confident in generalizing the findings produced from our sample of shows to the larger population of programs. This stands in contrast to the previous methodological design favored by most content-based studies, that of gathering a single intact calendar week of programming. That approach subjects the sample to potential biases that may systematically influence an entire group of programs, such as an upsurge in stories about love and sex during the week of Valentine's Day. The composite week sampling design was first developed for the National Television Violence Study (Wilson et al., 1997) and has been widely acknowledged as an important methodological innovation.

Channels in the Study
A total of ten channels were included in the study. These include the four major commercial broadcast networks (ABC, CBS, Fox, NBC), one independent broadcast station that is a WB network affiliate, one public broadcasting station (PBS), three basic cable channels (Lifetime, TNT, and USA Network), and one premium cable channel (HBO). The Los Angeles market was the site used for sampling all channels in the study during the previous three times of measurement (1997-98; 1999-2000; 2001-02). Beginning with this study of 2004-05 programming, sampling was shifted to the Tucson, Arizona market.

This change of sites for program sampling was carefully considered for any possible influence on the equivalence of the data for over-time comparisons. There is strong evidence to indicate that this shift poses no threat to the validity of the study based on careful scrutiny of the programming aired on the channels studied across these two markets.

For three of the four major broadcast networks, affiliates in each market air nationally distributed programming during the large majority of the time periods examined by this research. The handful of remaining hours are filled largely with local news coverage and syndicated programming such as *Dr. Phil*, *Oprah Winfrey,* and *Jeopardy*. Because our study excludes from examination daily or breaking news coverage, and because the nature of the syndicated programming delivered in major markets is highly similar from one locale to another (i.e., *Dr. Phil*, *Oprah Winfrey,* and *Jeopardy* air in virtually all major markets, though not necessarily on affiliate stations of the same network), the overall sample of programming gathered on these stations appears to vary little if any between Los Angeles and Tucson.

For the four cable networks studied (Lifetime, TNT, USA, and HBO), all programming is distributed nationally and thus site shifting poses no threat to the equivalence of the sample. While there is the potential for slight variability across markets for affiliates of Fox, PBS, and WB, in fact their affiliate stations demonstrate an extremely high degree of consistency in program offerings across these two markets. Even though choices are made at the local station level, the shows selected are all drawn from the same pool of nationally-available syndicated programming, and audience preference patterns apparently tend to drive similar program decision-making outcomes across major markets. Indeed, the high degree of consistency in programming on the channels in this study across these two locations serves to underscore our confidence in the broad generalizability of our findings to most major television markets.

One final consideration related to site shifting involves the time periods sampled for the study. The Los Angeles-based samples in the previous studies were gathered between the hours of 7:00 a.m. and 11:00 p.m. PST. To obtain equivalent programming in the Mountain Standard time zone, it was necessary to adjust the time periods for sampling forward one hour, starting at 6:00 am and ending at 10:00 p.m. MST. This reflects the standard network practice of airing the identical television programming one hour earlier in the country's two middle time zones, Central and Mountain, as compared to the East Coast and West Coast time zones. Thus, for example, the same prime-time programs that air between 8:00-11:00 p.m. on the East and West Coast are delivered between 7:00-10:00 p.m. in the Central and Mountain time zones. This shift in the hours sampled ensures the comparability of the current study to the previous data sets, and merely reflects normative variation in people's use of time in different regions of the country (Robinson & Godbey, 1997), which holds implications well beyond television viewing such as for waking and sleeping hours.

In sum, there is strong evidence to indicate that shifting the site of program sampling presents no threat to the comparability of the data and the integrity of the over-time comparisons for the study.

Program Eligibility

The composite week sampling design generates a representative collection of all programs presented on television across the channels studied. However, the design of this study excluded newscasts, sports, and children's programming from analysis. The news we have excluded is limited to programs identified as "daily news," which refers to coverage of time-sensitive, breaking events such as would be delivered on a nightly network newscast. Morning news/entertainment programs such as "Today" and prime-time news magazines such as "20/20 or "Dateline" were not classified as daily news, and thus were included in the analyses.

These exclusions are consistent with the orientation of this research, which is to assess the pattern of portrayals involving sexuality that are included in scripted entertainment programming. By excluding news, sports, and children's programs, we do not mean to suggest that any sexual information conveyed in these contexts is unimportant. Rather we have simply chosen to avoid diluting our focus by excluding the types of programming that present very different kinds of issues and concerns better addressed by a unique evaluation.

A total of 959 programs are included in the composite week sample analyzed for this research. Table 1 shows the breakdown of those programs for each of the channels in the study.

Table 1: Distribution of Programs Analyzed by Channel: Composite Week

CHANNEL	N
ABC	95
CBS	92
Fox	106
NBC	82
PBS	76
WB/Ind.	113
Lifetime	115
TNT	97
USA	100
HBO	83
TOTAL	**959**

Additional Sampling Details

The random selection sampling design may result in a small proportion of program overlap on each channel's composite week sample grid. This typically occurs near the end of the sampling period when only a small number of half-hour time blocks remain to be filled, and the programs aired during those periods are greater than a half-hour in length. All programs identified by the random selection process are always taped and included in the sample, and because a program can only be considered in its entirety there are some time blocks in which two programs rather than one were sampled on a particular channel. These are noted on the sample grids included in the appendices of the report.

The degree of program overlap in the sample is small, and does not present any threat to the generalizability of the study's findings. Quite the contrary, the independence of selecting shows individually and randomly provides significant strength to the generalizability of the findings, as compared to alternative sampling designs including the more common practice of taping a single calendar week of programming for analysis. Also, due to the nature of the sample design (i.e., programs selected for sampling are always taped and analyzed in their entirety), approximately 2% of the shows in the composite week either start before 6:00 a.m. or end beyond 10:00 p.m. MST. This occurs, for example, when a movie begins at 8:00 p.m. and does not conclude until 10:30 p.m. MST.

Prime-time Over-sample

As a complement to the composite week sample, we also collected a prime-time over-sample for the four major commercial broadcast networks (ABC, CBS, NBC, Fox) as indicated in Table 2. This set of shows consists of a total of three weeks of prime-time programming (7:00-10:00 p.m. MST) for each of the networks, or about 63 hours per channel (except for Fox, which delivers only two hours of network programming per night, rather than three). It was obtained using the same selection process of randomly sampling half-hour time blocks that was employed for gathering the composite week; likewise, it was assembled during the same sampling periods. A complete list of the programs included in the prime-time over-sample is included in Appendix B of this report.

It should be noted that prime-time network programming is still included in its proper proportion in the previously described composite week sample that represents the television environment as a whole. To more closely examine patterns in evening network programming, however, we supplemented the one week of prime-time material contained in the composite week with an additional two weeks worth of content, yielding a total of three weeks of programs for each channel. In most but not all cases, this design would yield three episodes of the same series. That does not result, however, for some time slots affected by program cancellations and/or series re-scheduling that occurs over the course of the television season. The prime-time over-sample includes a total of 261 shows. Of these, 87 are programs included in the composite week sample as well as the prime-time over-sample, while 174 additional programs are included solely in the prime-time over-sample.

Teenage Program Sample

As a complement to the two program samples described above, we also collect and analyze a sample of the programs most popular with teenage audiences. Three randomly selected episodes have been gathered for each of the 20 most heavily viewed television series for those 12-17 years old, regardless of network, as measured by the season cumulative totals we have obtained from the A.C. Nielsen Company. Table 3 provides a list of the Top 20 most frequently viewed programs by this age group.

In contrast to the composite week sample and prime-time over-sample, which have been gathered for all iterations of this ongoing series of studies, the teenage program sample was added to the project as part of the Sex on TV 3 report, which examined television programming from the 2001-02 season. Thus, longitudinal comparisons in this area are constrained to just two times of measurement at this point.

The sample of programs most frequently viewed by teens consists of a total of 60 programs which are analyzed and reported separately from the composite week sample and the prime-time over-sample. Of this total, 39 of the shows were obtained from programs in the composite week sample and prime-time over-sample, while 21 additional shows were randomly selected as necessary to record for analysis within this category.

Thus, from an overall perspective, the study includes a total of 959 programs in the composite week analysis, 174 additional shows that were part of the prime-time over-sample, and 21 unique programs that were added for the teenage program sample, yielding a grand total of 1154 programs examined for this research.

Table 2: Distribution of Programs Analyzed by Channel: Broadcast Network Prime-time

CHANNEL	N
ABC	69
CBS	69
Fox	54
NBC	69
TOTAL	261

Table 3: Top 20 Programs Most Frequently Viewed by Teens

PROGRAM	NIELSEN RATING
American Idol	9.7
Simpsons	5.0
Desperate Housewives	4.8
Survivor: Palau	4.4
CSI: Crime Scene Investigation	4.4
Extreme Makeover: Home Edition	4.4
The O.C	4.0
Family Guy	3.8
Survivor: Vanuatu	3.7
One Tree Hill	3.6
Nanny 911	3.2
Lost	3.2
That 70s Show	3.1
WWE: Smackdown	3.0
7th Heaven	2.9
Quintuplets	2.8
Without a Trace	2.8
24	2.8
Arrested Development	2.7
America's Next Top Model	2.7

Source: Nielsen broadcast and cable national audience estimates for the period 9/20/04 to 4/3/05 for 12-17 year-olds.

Note: Nielsen ratings list episodes from the same series that air at different and/or multiple times during the week as separate entries. For purposes of our top program list, only one entry was allowed per program title.

CONTENT MEASURES

This study performs scientific content analysis on the three groups of programs described above. In this section, we present the basic definitions we employ for identifying portrayals of sexual talk and behavior. We also explain the range of measures we apply to evaluate the contextual aspects of the portrayals identified in each of the areas of talk about sex and sexual behavior.

Levels of Analysis

Coding for any portrayal involving sexual content was performed at two distinct levels of analysis: the scene level and the program level. That is, some variables were measured solely on the basis of what happened within the scene in question, whereas others assessed broader contextual themes or issues that can only be judged at the end of a show, weighing all aspects of the program as a whole.

Scene level measures. The most basic and common unit of analysis for this study is the scene. A scene is defined as a sequence in which the place and time generally hold constant. Most scenes can be thought of in the same sense as a passage in a story; a scene ends when the primary setting shifts in time, place, or characters in a way that extensively interrupts the flow of related action. In our analysis, a commercial interruption always signals the end of a scene. Scenes are coded only when they are identified as containing sexual material according to the definition specified below.

Program level measures. While it is important to quantify the nature and context of sexual portrayals at the scene level, it is also important to consider the collective theme or pattern of messages a program conveys. The program level unit of analysis assesses broader thematic issues encompassing the program as a whole. Such judgments cannot necessarily be captured by simply adding up all of the more microscopic observations at the scene level, and thus we train coders to apply independent measures based upon everything they have seen throughout the entire show.

We now turn to the task of explicating our basic definitions and variables for analyzing sexual content.

Measuring Sexual Messages

For this study, sex is defined as any depiction of sexual activity, sexually suggestive behavior, or talk about sexuality or sexual activity. Portrayals involving only talk about sex are measured separately from those that include sexual actions or behaviors. Dialogue categorized as "talk toward sex" that often occurs concurrently with sexual behavior is not recorded to avoid double-coding.

Sexual dialogue, or what we term "talk about sex," involves a wide range of types of conversations that may involve first-hand discussion of sexual interests and topics with potential partners, as well as second-hand exchanges with others that convey information about one's prior, anticipated, or even desired future sexual activities. For purposes of measuring talk about sex, both the topic of reproductive issues (such as contraception or abortion) and sexually-transmitted diseases (including but not limited to HIV/AIDS) were considered as sexual.

To be considered a sexual behavior, actions must convey a sense of potential or likely sexual intimacy. For example, a kiss of greeting between two friends or relatives would not be coded as sexual behavior, whereas a passionate kiss between two characters with a discernible romantic interest would be. The lower threshold for sexual behaviors measured by the study was physical flirting, which refers to behavioral actions intended to arouse sexual interest in others, such as a woman licking her lips provocatively while gazing intently at a man in a bar. This example underscores that our measurement in this realm encompasses sexually-related behaviors, and should not be equated strictly with the consummate sexual behavior of intercourse. In addition, behaviors must be considered a substantial part of the scene in which they occur; portrayals which are judged as minor or peripheral (e.g., a couple of "extras" are shown "making out" in the background in a park scene which features two primary characters engaged in a serious non-sexual discussion) are not reported by the study.

Scene level contextual variables. Sexual dialogue, or what we term "talk about sex," involves a range of different types of conversations. We ultimately classified *type of talk about sex* into one of six distinct categories: comments about own/others' sexual actions/interests; talk about sexual intercourse that has already occurred; talk toward sex; talk about sex-related crimes; expert advice; and other. The first of these categories is by far the broadest, encompassing verbal exchanges about sexual relations that people wish they were having now, may want to have in the future, and so on. The second category involves comments about specific instances of sexual intercourse that have actually occurred, as distinct from what people want or try to promote. The third category, talk toward sex, involves efforts to promote sexual activity that are conveyed directly to the desired sexual partner. The fourth category, talk about sex-related crimes, involves any reference to illegal sex acts whether they have actually occurred, are simply feared, or are otherwise the subject of discussion. The fifth category, expert

advice, entails the seeking and delivering of sincere advice about sex from an authority figure, which is defined as someone who has received formal training relevant to the advice they deliver. Expert advice may occur in either a real setting, such as on a talk show, or in a fictional context, such as in a drama. Comments that met the definition for talk about sex indicated above but which did not fit any of the above categories were classified as "other."

The *type of sexual behavior* was measured using a range of six categories that began with physical flirting (behavior meant to arouse or promote sexual interest), and also included passionate kissing (kissing that conveys a sense of sexual intimacy), intimate touching (touching of another's body in a way that is meant to be sexually arousing), sexual intercourse strongly implied, and sexual intercourse depicted. Highly infrequent behaviors that meet the definition of sexual behavior indicated above but which do not fit in such any other category (e.g., self-gratification) were classified as "other."

The measurement of intercourse is particularly important, and the category termed "intercourse implied" is the only category of behavior in the study for which content is coded when the behavior is not shown overtly on the screen. Intercourse implied is said to occur when a program portrays one or more scenes immediately adjacent (considering both place and time) to an act of sexual intercourse that is clearly inferred by narrative device. Common examples would include a couple kissing, groping, and undressing one another as they stumble into a darkened bedroom, with the scene dissolving before the actual act of intercourse ensues; or a couple shown awakening in bed together with their conversation centering on the lovemaking they had performed before falling asleep. Such portrayals are not necessarily explicit in any way but clearly convey the message that sex has occurred, and thus it is essential that such portrayals be reflected in our content measures.

In contrast, "intercourse depicted" is judged to occur when a direct view is shown of any person who is engaged in the act of intercourse, regardless of the degree of nudity or explicitness presented. Discreet portrayals may show a couple only from the shoulders up when they are engaged in intercourse. As we explain below, the explicitness of any sexual behavior is measured independently of the judgment about the type of behavior that occurs.

For any material involving either sexual dialogue or behavior, the degree of *scene focus on sex* is judged, differentiating minor or inconsequential references and depictions from portrayals in which there is a substantial or primary emphasis on sex. In addition, all scenes that include sexual behavior are coded for *degree of explicitness*, which indicates the physical appearance of the characters involved in the behavior. The categories for coding included provocative/suggestive dress or appearance (attire alone reflects a strong effort to flaunt one's sexuality); characters begin disrobing (the removing of clothing that reveals parts of the body not normally exposed); discreet nudity (characters are known to be nude but no private parts of the body are shown); and nudity (baring of normally private parts, such as the buttocks or a woman's breasts).

Finally, when a scene includes sexual content, coders also determine whether that scene includes any mention or depiction of *sexual risks or responsibilities.* This term is used to describe the issues surrounding the serious outcomes that can be associated with human sexual activity. In applied terms, sexual risks or responsibilities refer to such concerns as unwanted pregnancy or sexually transmitted diseases, and are described in greater detail immediately below in the section explicating program-level measures. Sub-categories employed to identify different types of sexual risks or responsibility messages include mention or use of a condom or other contraception; mention of "safe sex;" concern about or depiction of HIV/AIDS, STDs, unwanted pregnancy, or abortion; and mention or depiction of abstinence or waiting for sex.

For content judged to fit within any of these categories, the coder also evaluated several other aspects of the scene. The first of these assessed whether the *scene focus* on sexual risks or responsibilities was primary, substantial, minor, or inconsequential. Another variable identified the type of character (e.g., parent, peer) who was the *source of information* for the sexual risk or responsibility message. And a third contextual measure in this realm examined the *valence* associated with the presentation of the risk or responsibility message, taking into account the scene as a whole. Coding options for this variable included primarily positive (reflecting support for or concern about sexual risk/responsibility issues), primarily negative (minimizing concern for sexual risk/responsibility issues), mixed, or neutral/can't tell.

Program level variable. To complement the scene level variables, an assessment was conducted at the overall program level judging whether or not each show that contains any sexual content places strong emphasis throughout on a *risks or responsibilities program theme.* Three distinct risk or responsibility program themes are examined: (1) sexual patience; (2) sexual precaution; and (3) depiction of risks and/or negative consequences of sexual behavior.

The first of these themes, sexual patience, encompasses programs that place emphasis on abstinence from sex or waiting for sex as either a positive moral stance or as a sound approach to avoiding the risks of STDs or unwanted pregnancy. The second theme, sexual precaution, refers to the use or discussion of preventative measures (e.g., condoms) to reduce the risk of STDs or unwanted pregnancy. The third theme, depiction of risks, involves emphasis on the life-altering (e.g., unwanted pregnancy) or life-threatening (e.g., transmission of HIV/AIDS) outcomes that may result from unplanned and/or unprotected sexual intercourse. Across all of these areas, the applicable theme must be central to the program plot to be coded as an overall theme of sexual risk or responsibility.

Variable scaling information. In the analyses we employ to generate findings for the study, some of the individual variables described above have been combined to create an index or scaled in a way that will help to simplify the presentation of data. Here we provide information that explains how we have calculated several basic measures that we present in our subsequent report of findings.

To assess the level of sexual behavior, we report values on a scale of 1 to 4: a value of 1 indicates physical flirting, a value of 2 indicates intimate touching or passionate kissing, a value of 3 reflects sexual intercourse strongly implied, and a value of 4 represents intercourse depicted. Explicitness is measured on a scale of 0 to 4, with 0 indicating none, 1 indicating suggestive/provocative dress, 2 indicating disrobing, 3 reflecting discreet nudity, and 4 indicating nudity. Both of these scales are reported as a threshold score within each scene. For example, a scene that contains kissing and intercourse strongly implied yields a score of 3, the higher of the two behavior values. Similarly, a scene in which disrobing occurs followed by discreet nudity is recorded as a 3.

To assess the level of talk about sex, we are constrained by the fact that there is no apparent rationale for assigning greater or lesser values to any one of the various categories of sexual dialogue over another for purposes of considering their implications for audience effects. Similarly, there is no obvious validity for assigning greater weight to scenes that involve several such categories (e.g., talk about one's interest in sex, and talk about sexual intercourse that has occurred) rather than just a single one, as one scene could involve elaborate sexual discussion within one category while another scene could encompass two categories of talk but treat both superficially.

Given these limitations, we have chosen to construct a scale for the level of talk about sex that considers all scenes that present differing categories of dialogue as being of the same potential weight; and we have then based our calculation on the judgment that indicates the degree of focus, or emphasis, placed on any applicable talk category within the scene. The degree of focus involving talk about sex was measured on a four point scale reflecting a continuum from minor to primary emphasis within each scene. Of the available options, we believe that the degree of focus is the best estimate of the meaningfulness and potential impact of the talk, and thus we have grounded our measurement for talk about sex in it.

Content Coding and Reliability

This section reports the process employed to review and evaluate the program samples to obtain data for the study. The scientific integrity of the content analysis data reported in this research is established in large part by careful statistical monitoring of the inter-coder reliability of judgments. That process started well before any actual coding of data was performed.

A group of 17 undergraduate students at the University of Arizona served as coders for this project. Coders were trained approximately six hours per week over a 15-week period to apply the full range of measures designed for the study, which are detailed in an elaborate codebook of rules. The training process included extensive practice in a viewing lab, with each coder's performance monitored systematically to diagnose any inconsistencies in

their interpretation and/or application of the content measures. At the conclusion of training, a statistical test of inter-coder reliability was conducted to verify the strength of the consistency of their judgments. The results of the final training test are reported below alongside the findings for the reliability assessments performed during the actual process of data collection.

Once training was complete, the coding of data was accomplished by randomly assigning individual coders to view programs and to apply our content measures. Coders viewed each show alone in a video lab and were allowed to watch any given segment as many times as necessary to correctly apply the measures. Data for each program were obtained from a single coder. For this reason, it is necessary to demonstrate that the coding process maintained a strong and consistent level of reliability over time in order to ensure the quality of the data.

Assessing the Reliability of the Data
The coding process required approximately 12 weeks to complete. To assess the reliability among the coders as they were performing their work, a randomly selected program within a specified genre of content was independently evaluated by all coders. This process was repeated at regular intervals spaced approximately 10 days apart during the period when the coding work was being accomplished. For each reliability test, the coding judgments on a single program were then compared across all coders for reliability assessment purposes.

Conceptualization of Reliability
Coders must make a variety of different types of decisions when viewing a show. These decisions exist at two distinct levels. The first focuses on unitizing, or the identification of scenes containing any sexual content. At this level, a coder is watching solely to determine whether the material meets the basic definition for sex. In addition, once coders identify a scene as containing sexual content, we must examine their consistency in classifying the portrayals within the scene.

In the sections that follow, we detail the specific procedures employed to calculate inter-coder reliability. This process is patterned after the approach devised for the National Television Violence Study (see Wilson et al., 1997), which describes the development of the procedures in greater detail. This approach reflects the most current methodological innovation for calculating reliability across large numbers of coders who are rendering content-based judgments at multiple levels of analysis (Potter et al., 1998). It involves independent assessment first of the fundamental unitizing judgments, followed by a discrete examination of the contextual measures that apply once the higher order units of analysis have been established.

<u>Agreement on unitizing.</u> Unitizing refers to the process of identifying each scene that contains any sexual content. Every time a coder identifies a scene with some codable material, s/he creates a line of data that includes a string of values indicating judgments for each applicable contextual variable. In evaluating the unitizing process, the focus is not on the agreement of the values for the contextual variables; rather, the aim is to assess the extent of agreement that a given scene contained sexual content.

In assessing reliability, if all coders identify the same number of scenes on their coding form for a show and if those scenes refer to the same scenes from the program, then there is perfect agreement. Both conditions must be met for perfect agreement to occur. If coders differ on the number of scenes identified, then there is not perfect agreement. If coders all have the same number of scenes, but there is disagreement about the scenes that were coded, then there also is not perfect agreement.

Three descriptors are reported for unitizing: the agreement mode, the range of scenes, and a statistic called the Close Interval around the Agreement Mode (CIAM). An example will explain what is meant by "agreement mode." If there are ten coders and one reported 9 scenes with sex, eight reported 10 scenes, and one reported 11 scenes, the mode would be 10 scenes as this is the number reported by the greatest number of coders. Thus, 80% of the coders are at this mode. Recall, however, that coders must identify the same scenes in order to have agreement. If all eight coders identified the same 8 scenes, then the agreement mode is 8.

Coders have to make many difficult judgments as part of the coding process. As a result, not every coder is at the agreement mode for every program, so we also report the range of scenes identified by the set of coders for each reliability test. The smaller the range, the tighter the pattern of agreement. However, the range can

sometimes be misleading as an indicator of the degree of variation in a distribution. For example, consider a case where there are ten coders and one identifies 4 scenes with sex, eight indicate 5 scenes, and one identifies 8 scenes. The range reported would be from 4 to 8 scenes, which appears to signal a wide range of disagreement. That interpretation would be inaccurate, however, as 90% of the coders are actually within one scene of the mode.

The most important statistic for evaluating reliability at this level is the Close Interval around the Agreement Mode (CIAM). We operationalize "close to the agreement mode" as those judgments that are within one scene (or 20% as described below) of the modal judgment. Thus, if the agreement mode for a program was 5 scenes of sex, we would include in the CIAM each of the following: (a) all coders who identified all 5 of the same scenes; (b) all coders who also saw 5 scenes but disagreed on just one of the scenes identified by those in the modal group; (c) all coders who saw only 4 scenes but each of those scenes matched the 5 scenes identified by the modal group; and (d) all coders who reported 6 scenes where 5 of those scenes were identical to the ones identified by the modal group. When the agreement mode is ten or greater, we establish the width of the CIAM as 20% on either side of the mode. For example, if the agreement mode is 10, we include coders who exhibit no more than two disagreements with the coders at the agreement mode.

Agreement on the contextual variables. The other important aspect of reliability is the degree of consistency among coders in choosing values for each contextual variable once they have identified the examples of sexual content. For program level measures, reliability was assessed by identifying the modal value for all coders. Percentage of agreement was computed by dividing the number of coders at the modal value by the total number of coders.

For scene level measures, it was necessary to construct a matrix for each of the context variables. For each variable, a column is entered for every coder, and a row for every scene that was identified by one or more coders as containing some codable portrayal in that area (i.e., talk about sex or sexual behavior). Each row of the matrix is then examined for its modal value for each applicable contextual variable. Next, the number of coders at the modal value is summed and entered as a marginal. The marginal totals are summed down across all scenes in the matrix for the same variable. This sum of the marginals (i.e., agreements) is then divided by the total number of decisions reflected in the entire matrix (i.e., all agreements and disagreements), and the resulting fraction yields the percentage of agreement among coders on that variable.

While the operational details are intricate, the concept of reliability is not. The term "percentage of agreement" simply refers to the number of times coders actually agreed, divided by the number of times they could possibly have agreed. The larger the result, the better the agreement.

Results of Reliability Testing
The mean agreement for identifying scenes that contained sexual content across all programs was 89% on the CIAM measure (see Table 4). The degree of consistency for unitizing, or identifying both sexual behavior and sexual dialogue within scenes, is highly credible given the complexity of the task and the number of coders involved. The consistency for coding the scene-level contextual variables was also very strong, achieving agreement at 90% or above on 26 of the 29 measures reported in the study. Inter-coder reliability on the overall program-level theme variable was 96%. Across all measures, no variable obtained a reliability coefficient below 83%, and thus all data for the study can be interpreted with confidence.

To summarize, tests to assess the degree of inter-coder agreement were performed throughout all phases of the data collection process. These tests demonstrate that the content measures applied in the study yielded highly reliable data from the coders who were reviewing the programming. Overall, the reliability analyses establish strong confidence in the quality of the data reported in the study.

Table 4: Reliability for Sexual Dialogue and Sexual Behavior Measures

	Unitizing								
Measures	**Grounded for Life**	**Boston Legal**	**Survivor: Vanuatu**	**General Hospital**	**Two Weeks Notice**	**Primetime Live**	**Malcolm in the Middle**	**House**	**Overall Means**
Scene Range	8-15	8-12	0	4-6	12-17	1-5	7-10	2-3	
Scene Mode	10	11	0	6	14	2	9	3	
CIAM	88%	92%	100%	80%	79%	94%	81%	100%	**89%**

Scene Level Context Variables

Talk About Sex									
Own/Others	88%	89%	100%	89%	96%	94%	82%	90%	**91%**
Talk About	99%	96%	100%	99%	97%	100%	100%	100%	**99%**
Talk About First Time	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
Talk Toward	99%	95%	100%	100%	100%	100%	98%	100%	**99%**
Talk about Sex Crimes	91%	99%	100%	99%	100%	100%	100%	95%	**98%**
Expert Advice	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
Other	100%	98%	100%	94%	99%	83%	100%	100%	**97%**
Oral Sex	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
Talk Focus	78%	85%	n/a	99%	84%	78%	76%	81%	**83%**
Talk Character	98%	100%	n/a	99%	99%	100%	98%	98%	**99%**

Behaviors									
Flirtatious Behavior	97%	95%	100%	100%	96%	100%	96%	95%	**97%**
Kissing	100%	97%	100%	100%	100%	100%	100%	100%	**99%**
Intimate Touch	100%	99%	100%	92%	100%	100%	93%	90%	**97%**
Intercourse Implied	100%	99%	100%	91%	100%	100%	99%	100%	**99%**
Intercourse Depicted	100%	99%	100%	97%	100%	100%	100%	100%	**99%**
Other	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
Oral Sex Implied	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
Oral Sex Depicted	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
Behavior Character	94%	100%	n/a	100%	100%	n/a	99%	93%	**98%**
Explicitness	76%	100%	n/a	100%	94%	n/a	97%	80%	**91%**

Risk/Responsibility									
Topic	99%	99%	100%	99%	100%	100%	100%	100%	**99%**
Focus in Scene	79%	81%	n/a	96%	100%	n/a	n/a	n/a	**89%**
Source of Information	95%	100%	n/a	98%	100%	n/a	n/a	n/a	**98%**
R/R Character	99%	100%	n/a	97%	100%	n/a	n/a	n/a	**99%**
Valence	76%	88%	n/a	84%	100%	n/a	n/a	n/a	**83%**

Special Intercourse Measures

First Time	n/a	100%	n/a	100%	n/a	n/a	100%	n/a	**100%**
Relationship	n/a	75%	n/a	87%	n/a	n/a	100%	n/a	**87%**
Presence of Drugs	n/a	100%	n/a	100%	n/a	n/a	100%	n/a	**100%**
Presence of Alcohol	n/a	92%	n/a	100%	n/a	n/a	100%	n/a	**97%**

Program Level Context Variable

Program Theme	89%	86%	n/a	100%	100%	100%	100%	100%	**96%**